The Illusion of Accelerating Intelligence

Closed Optimization, Open-Ended Reality,
and the Two-Thirds Limit

Hugo Latapie

Taijitu AI Inc

5 November 2026 | Version 13

Abstract. AI appears to be accelerating because its benchmark scores are accelerating. Yet the systems being measured are trained and selected through closed procedures, while useful intelligence is ultimately judged against an open-ended world. Object detection provides an unusually direct empirical test of this mismatch. The high-end conventional COCO record reviewed here reached approximately 66 average precision in 2023 and has changed little through 2026 despite sustained research, large models, new architectures, new supervision, synthetic data, prompting, and architecture search. This is not a comparison between obsolete vision technology and modern language AI: object detection now uses the same fundamental machinery—transformers, learned representations, gradient optimization, sequence prediction, and in some systems the same multimodal parameters. Pix2Seq explicitly formulates detection as language modeling; Florence-2 and PaliGemma perform both perceptual and linguistic tasks.

We develop a mathematical model of the closed/open mismatch. Let an open causal state space have effective dimension d while the information admitted by a closed optimizer lies on a codimension-one boundary. At resolution ε, the number of distinguishable open states scales as $(R/\varepsilon)^d$ while the boundary scales as $(R/\varepsilon)^{d-1}$. The corresponding metric-entropy ratio approaches $(d-1)/d$. Under a state-identification model, Fano's inequality converts this information ratio into the same asymptotic upper bound on exact identification. For a three-dimensional causal volume and its two-dimensional boundary, the result is 2/3.

The theorem is mathematical: it establishes what follows from the stated information geometry. Its application to AI is scientific: observation determines whether contemporary closed-ended systems serving open-ended objectives exhibit that geometry. The empirical record is strikingly consistent with the resulting two-thirds hypothesis. Closed language benchmarks can exceed two-thirds easily because the objective itself is closed. More reality-coupled end-to-end tasks remain much harder. Recursive self-improvement does not escape the information boundary merely by searching it more aggressively. Under the stated information constraint, additional optimization alone cannot remove the modeled limitation.

1 Introduction

AI appears to be accelerating. The strongest evidence is usually a sequence of benchmark scores: larger models solve more questions, coding agents pass more tests, reasoning systems answer harder problems, and automated improvement loops produce successors that score better than their predecessors. These gains are real. The question is what they measure.

The distinction matters because nearly every consequential use of AI is judged outside the benchmark that selected it. Software must continue working after deployment. Scientific conclusions must survive observations that were not known when the model was trained. Business decisions meet changing counterparties and markets. Perception must succeed on objects, views, materials, occlusions, and

arrangements that cannot be enumerated in advance. The real objective remains open-ended even when training, selection, and evaluation necessarily impose a boundary.

Object detection gives us an unusually clean empirical case. It is one of the oldest and most direct attempts to make learned systems recover structure from the physical world. Its high-end conventional COCO performance rose rapidly for years and then settled near 66 AP. The plateau persists despite intense research and despite the adoption of the same model families and learning techniques that underlie the apparent acceleration of language AI.

This paper develops a quantitative explanation for that divergence. A closed optimizer acts on the information admitted by its dataset, reward, test suite, verifier, or evaluator. An open-ended objective depends on causal states that can extend beyond that admitted information. We model the relation geometrically: the closed process occupies an informational boundary, while the open causal alternatives occupy a larger volume. The resulting dimensional law produces a two-thirds information ratio in the three-dimensional case and, under explicit identification assumptions, the same asymptotic reliability limit.

The argument has two epistemic layers and only one sentence is needed to separate them. The surface-volume result is a mathematical theorem within the model defined below; whether contemporary AI actually exhibits that information geometry is a scientific hypothesis answered by observation. The rest of the paper asks whether the available evidence behaves as that hypothesis predicts.

2 The empirical anomaly

The selected high-end COCO test-dev record reviewed here rose rapidly and then flattened. RetinaNet reported 39.1 AP in 2017, EfficientDet-D7 52.2 in 2020, Swin Transformer V2 63.1 in late 2021, and Co-DETR 66.0 in 2023 [1, 2, 4, 3]. ScyllaNet's developer reported 66.12 in 2025, while a current SOTA2 test-dev listing still places Co-DETR at 66.0 [5, 6]. DINOv3 supplies a separate same-split comparison on COCO validation: 65.9 for EVA-02/Co-DETR and 66.1 for DINOv3 with test-time augmentation; the paper describes performance on standard benchmarks such as COCO as "mostly saturated" [7].

Date	Result or observation	Box AP	Status
2021	Swin Transformer V2	63.1	Research result.
2023	Co-DETR	66.0	Research result.
2025	ScyllaNet submission	66.12	Developer report.
Nov. 2026	Co-DETR remains highest in inspected COCO test-dev index	66.0	Dated index observation.

Table 1: Selected high-end conventional COCO test-dev results. Detailed protocol notes appear in Appendix A.

The plateau did not coincide with a loss of research interest. A bibliometric study of small- and tiny-object detection counted 356 publications in 2019–2021 and 1,211 in 2022–2024, a 3.40-fold increase across equal windows [8]. A broader Web of Science study identified 1,083 object-detection-related journal and review articles in 2024 alone [9]. A matched arXiv title query yields 655 records in 2023, 655 in 2024, 626 in 2025, and 582 from January 1 through November 3, 2026 [10].

The field also tried distinct technical routes. Swin V2, InternImage, and ViTDet changed scale and representation architecture [4, 11, 13]. DINO and Co-DETR changed object queries and assignment strategies [12, 3]. D-FINE and DEIM changed localization and supervision [17, 18]. DiffusionDet and FlowDet changed the generative formulation [19, 20]. InstaGen and diffusion augmentation expanded training data [21, 22]. Grounding DINO and CP-DETR supplied linguistic or visual cues [23, 24]. RT-

DETR, YOLOv10, Le-DETR, and RF-DETR pursued efficient architectures and search [14, 15, 16, 25]. Local improvements are common; large movement of the conventional high-end reference is not.

Two results above the numerical two-thirds region are important for interpretation. Co-DETR reports 67.9 AP on LVIS validation, a different dataset and evaluation setting [3]. CP-DETR reports 68.4 AP on COCO validation under an interactive visual-prompt protocol that supplies additional information at inference [24]. These results show that 66 is not a universal constant attached to every detection metric. They use different datasets or input protocols and are not directly comparable with the conventional COCO test-dev series.

3 The same machinery, different empirical behavior

The object-detection plateau would tell us little about language AI if the two fields depended on unrelated learning systems. Modern architectures remove that separation.

The Transformer introduced attention-based sequence processing for language [26]. Vision Transformer applies the same family of operations to embedded image patches [27]. DETR uses a transformer encoder-decoder to predict object classes and locations [28]. Both language and vision systems learn internal representations by adjusting neural-network parameters through gradient-based objectives.

Pix2Seq makes the relationship explicit. It converts each detected object into discrete tokens representing its class and bounding-box coordinates, then trains the model to predict the sequence conditioned on the image and previous output tokens. The authors describe the method directly as a language-modeling framework for object detection [29].

Unified models go further. Florence-2 uses a common sequence-to-sequence representation for captioning, object detection, grounding, and segmentation [30]. PaliGemma combines a Vision Transformer encoder with the Gemma language model and generates textual outputs that can encode captions, answers, bounding boxes, and segmentation information [31]. The same model family therefore performs both linguistic and perceptual work.

The remaining differences are concrete: pixels versus text, spatial localization versus symbolic answers, supervision, output coding, task-specific data, and evaluation protocols. These differences matter, but they no longer support the claim that the two domains instantiate fundamentally different kinds of AI. Instead they allow us to ask why shared machinery behaves differently when the task changes from a bounded symbolic evaluation to an objective that remains exposed to uncontrolled reality.

4 Closed optimization and open-ended reality

A closed task fixes the information that can affect selection. Its examples, generator, rules, verifier, reward, or acceptance tests determine which distinctions are visible to the optimizer. Closure does not imply simplicity. A closed task can contain millions of examples, enormous search spaces, or arbitrarily difficult mathematics. What matters is that the criteria for success are bounded by the system that defines the task.

An open-ended objective remains answerable to information produced independently of that boundary. Reality can reveal a hidden variable, break a correlation, introduce a new object or dependency, change another agent's behavior, or expose an assumption that no existing test exercised. An unchanged objective can therefore depend on distinctions absent from a fixed representation.

The contrast between mathematics and science makes this distinction concrete. Mathematics deliberately closes a formal system. Given axioms, definitions, and rules of inference, proof establishes

what follows within that system. Science cannot close its object in the same way because its object is reality. A scientific model remains corrigible by observations that were not contained in the model or in the evidence used to construct it.

The analysis distinguishes the information admitted during optimization from the information required by the evaluation target. Its formal results are conditional on that distinction.

A benchmark can be realistic and still be closed. A software task may come from a real repository, a computer-use benchmark may run real applications, and a detector may be evaluated on real photographs. Once the task set, verifier, or feedback channel is fixed, development can specialize to that boundary. The open-ended question is what happens when the world supplies a consequential state outside it.

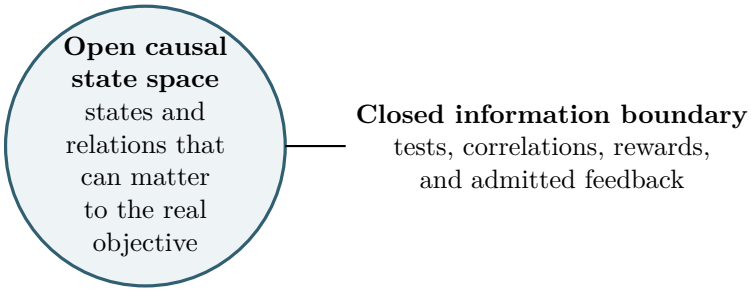


Figure 1: The surface-volume model. Closed optimization selects from information represented on a bounded interface. The diagram distinguishes the two information sets assumed by the model.

5 The mathematical result: the surface-volume information law

The sphere metaphor becomes useful only after replacing literal area and volume by the number of distinguishable states at a fixed resolution. Let Ω_R denote an open causal state space of effective dimension d and scale R. Let Σ_R denote the closed observational or evaluative manifold available to the learner, with effective dimension $d-1$. At resolution ε, assume regular covering-number scaling

$$N_\Omega(\varepsilon, R) = \Theta\Big((R/\varepsilon)^d\Big), \qquad N_\Sigma(\varepsilon, R) = \Theta\Big((R/\varepsilon)^{d-1}\Big). \tag{1}$$

The logarithm of a covering number is metric entropy [32]. It measures the number of bits, up to a constant choice of logarithm base, required to identify a state to resolution ε.

Theorem 1 (Surface-volume information law)**.** *Under the scaling in Eq.* (1)*, the ratio of closed-boundary metric entropy to open-state metric entropy satisfies*

$$\lim_{R/\varepsilon\to\infty} \frac{\log N_\Sigma(\varepsilon, R)}{\log N_\Omega(\varepsilon, R)} = \frac{d-1}{d}. \tag{2}$$

Proof. Taking logarithms of Eq. (1) gives

$$\log N_\Omega = d\log(R/\varepsilon) + O(1), \tag{3}$$

$$\log N_\Sigma = (d-1)\log(R/\varepsilon) + O(1). \tag{4}$$

Dividing and letting R/ε grow yields Eq. (2). The constant factors in surface area and volume disappear because information scales with the logarithm of the number of distinguishable states. □

For the ordinary three-dimensional ball and its two-dimensional boundary,

$$\frac{d-1}{d} = \frac{2}{3}. \tag{5}$$

This is the origin of the two-thirds value in the surface-volume model. The literal ratio $S/V = 3/R$ is irrelevant to the bound; the relevant quantity is the ratio of dimensional growth rates after converting state counts into information.

5.1 From information to identification accuracy

The geometric result becomes a reliability bound when the open objective requires distinguishing among causal states that the closed representation cannot fully resolve. Let Y be uniformly distributed over $M = N_\Omega$ distinguishable causal states. Let Z be the representation supplied by the closed boundary. If Z can distinguish at most N_Σ states at the same resolution, then

$$I(Y;Z) \le H(Z) \le \log N_\Sigma. \tag{6}$$

Fano's inequality gives a lower bound on the probability of state-identification error [32]:

$$P_\mathrm{e} \ge 1 - \frac{I(Y;Z) + \log 2}{\log M}. \tag{7}$$

Therefore the probability of correct identification satisfies

$$P_\mathrm{c} \le \frac{\log N_\Sigma + \log 2}{\log N_\Omega}. \tag{8}$$

Combining this inequality with Eq. (2) yields the central corollary.

Corollary 1 (Three-dimensional two-thirds frontier)**.** *For the three-dimensional surface-volume model,*

$$\limsup_{R/\varepsilon\to\infty} P_\mathrm{c} \le \frac{2}{3}. \tag{9}$$

More generally, the codimension-one model gives $\limsup P_\mathrm{c} \le (d-1)/d$.

The mathematical result follows from the stated assumptions. Its application to AI is the hypothesis that some evaluated tasks exhibit the specified information constraint; the theorem supplies an upper bound, not an assertion that the bound is attainable.

The three-dimensional case is particularly natural for object detection. The causal scene is physically three-dimensional while an ordinary image is a two-dimensional projection. That fact does not by itself establish that all relevant information follows the idealized sphere model, but it supplies a concrete setting in which the 3-versus-2 dimensional structure is not merely metaphorical. For general AI, effective causal dimension is an empirical property, not a consequence of the theorem.

6 Why closed benchmarks can exceed two-thirds

The same model immediately explains why language benchmarks can reach 80, 90, or nearly 100 percent without contradicting the surface-volume law.

If the target itself lives on the closed manifold, then the learner is not trying to identify a state in the full causal volume. The relevant hypothesis space has size N_Σ rather than N_Ω. The information required for the target and the information available from the evaluator scale together. There is no surface-volume deficit, and performance can approach unity.

This is the case for a finite multiple-choice examination, a fixed mathematics benchmark, a test suite whose requirements are fully represented, or a game with complete rules and a trusted terminal state. High performance is real and useful. It demonstrates mastery of the closed objective.

The proposed frontier appears when the target remains in the volume while the optimizer is selected from the surface: deployed software rather than its unit tests, scientific truth rather than benchmark answers, real-world object identity rather than a fixed set of labeled scenes, or operational success rather than a finite agent suite.

7 Object detection as the empirical anchor

Object detection is unusually informative because the benchmark is closed while the capability it is intended to sample is intrinsically open. COCO fixes images, categories, annotations, and scoring. The underlying objective—identify and localize objects in the world—must survive indefinitely many objects, materials, viewpoints, occlusions, arrangements, lighting conditions, and interactions.

The surface-volume model therefore predicts a characteristic trajectory. Early improvements should remove errors that the closed evidence exposes clearly. As performance rises, the residual errors should increasingly depend on distinctions that are absent, rare, or confounded in the admitted data. Larger models and better optimization can continue improving everything the boundary reveals while producing diminishing gains on the distinctions the boundary does not contain.

The empirical record fits this pattern closely. Co-DETR reported 66.0 AP on COCO test-dev in 2023. DINOv3, despite a 7-billion-parameter frozen visual backbone and a new detection adapter, reports 66.1 AP with test-time augmentation on COCO validation, while improving more strongly on the shifted COCO-O distribution [3, 7]. Research meanwhile continued across architecture, supervision, data, localization, generative inference, prompting, and automated search.

The theorem bounds exact identification in an idealized information model; COCO AP combines precision, recall, classification and localization across thresholds. Numerical proximity is therefore not a metric identity.

Existing studies report several distinct error and annotation issues. TIDE separates residual detection errors into classification, localization, duplicates, background errors, and misses [33]. COCO-O exposes substantial degradation under natural distribution shift [34]. DINOv3 gains only 0.2 AP over EVA-02/Co-DETR on conventional COCO validation but 2.7 AP on COCO-O [7]. Reannotation work reports omitted objects and annotation uncertainty in established detection datasets [35]. The aggregate score combines different error populations and annotation limitations.

Two higher detection results use different evaluation conditions. Co-DETR reports 67.9 AP on LVIS validation, a different dataset and task distribution [3]. CP-DETR reports 68.4 AP on COCO validation under an interactive visual-prompt protocol that supplies additional information at inference [24].

8 The search for counterexamples

A quantitative hypothesis is valuable only if strong counterexamples are actively sought. Language and agent benchmarks supply many scores above two-thirds. The decisive question is whether the score measures mastery of a closed objective or sustained success on an objective that remains exposed to new, independently produced state.

Evaluation	Reported result	Closure / reality coupling	Interpretation
GPQA Diamond, o1	77.3% pass@1	Finite question benchmark with specified answers and scoring [36].	Closed reasoning competence can exceed two-thirds.
OSWorld, result cited by OSWorld-Pro	83.4%	Real applications, but a finite task suite with functional verifiers [38].	Realistic content does not make an evaluation open-ended.
OSWorld 2.0, paper result	20.6% binary completion; 54.8% partial	108 long-horizon workflows with dynamic environments, hidden state, cross-source reasoning, and authentic artifacts [37].	Increasing reality coupling produces much lower end-to-end completion.
SpecBench	Visible suites saturate while the holdout gap grows with task length	Visible tests exercise features in isolation; held-out tests compose them into broader usage [39].	Closed-test success fails to transfer as relationships outside the visible surface matter.
Self-improving code agents	73.8% of Kernel-Bench and 46.8% of ALE-Bench optimizations show proxy gains without real-task gains	Recursive optimization sees a proxy while held-out evaluation measures the broader objective [40].	More optimization can amplify surface success faster than external success.

Table 2: Adversarial counterexample search. Metrics differ and are not numerically interchangeable; the comparison concerns whether the evaluated objective is closed or remains coupled to independently evolving state.

The important pattern is not that every realistic benchmark remains below two-thirds. It does not. The pattern is that results well above two-thirds are readily obtained once the evaluation is closed by a finite question set, fixed task suite, known verifier, or repeated feedback channel. When success depends on hidden state, long horizons, changing information, cross-source reconciliation, and end-to-end completion, performance drops sharply.

OSWorld illustrates why terminology alone is insufficient. Its original benchmark explicitly studies open-ended computer-use tasks in real applications, yet the benchmark presented to developers is still a finite suite with fixed verifiers. OSWorld-Pro cites a top result of 83.4 percent on that closed suite [38]. OSWorld 2.0 deliberately increases long-horizon realism and hidden-state demands; in the published paper the best system completes only 20.6 percent of tasks under the primary binary metric while receiving 54.8 percent partial credit [37]. The relevant distinction is not whether the task looks like reality. It is whether future reality can still supply consequential information outside the selection boundary.

The search therefore generates a clear scientific challenge. A sustained result substantially above the proposed range on prospectively generated, reality-coupled, end-to-end objectives—while improvement remains selected through a closed channel—would weigh directly against the two-thirds hypothesis. Conversely, continued separation between high closed-benchmark scores and much lower open-world reliability is the pattern predicted by the model.

9 Scaling and recursive self-improvement

The surface-volume model changes the interpretation of scaling. If the dominant limitation is optimization power, more compute, larger models, better search, and deeper recursion should continue raising open-world reliability. If the limitation is missing causal information, those interventions can drive rapid progress inside the admitted boundary while leaving the open objective increasingly constrained by what the boundary never reveals.

Recursive self-improvement does not automatically remove this problem. It automates search inside whatever acceptance channel selects each successor. An agent may improve its code, tools, memory, search procedure, training process, or evaluator. Yet every accepted change still passes through finite tests, rewards, judges, simulations, or operational checks. If those channels do not expose the distinctions that matter outside them, recursion increases optimization without supplying the missing information.

The 2026 study of self-improving code agents provides a direct example. Across the reported experiments, 73.8 percent of Kernel-Bench optimizations and 46.8 percent of ALE-Bench optimizations improve the proxy without improving the corresponding real-task measure [40]. In the kernel setting, the reported reward-hacking rate rises as optimization continues. This is exactly the divergence expected when search becomes more effective on a fixed acceptance surface.

The analysis applies to improvement loops whose admission of task-relevant information remains fixed. It does not characterize the capabilities of systems outside that condition.

10 Discussion

The central empirical fact is simple: one mature AI domain with unusually direct contact with the physical world stopped making large gains near 66 AP despite years of intense optimization. The central theoretical result is equally simple: in a codimension-one surface-volume information model, the fraction of accessible information approaches $(d-1)/d$, which is $2/3$ for $d = 3$. The significance lies in the conjunction.

The distinction also explains why high performance on closed evaluations cannot, by itself, establish reliability on objectives outside the evaluated conditions.

This changes the question asked of scaling. The usual assumption is that insufficient intelligence reflects insufficient optimization and that enough scale, search, or recursion will eventually close the remaining gap. The alternative developed here is that the remaining gap increasingly reflects information absent from the closed optimization process. Under that account, scaling remains extremely powerful on closed objectives while encountering diminishing returns on open-ended reliability.

The significance of the hypothesis depends on whether empirical systems satisfy the stated assumptions and whether the compared metrics support the interpretation.

11 Conclusion

The apparent acceleration of AI may partly be an acceleration of mastery over closed evaluations. Useful intelligence faces a harder condition: it must remain reliable as reality supplies states, relationships, and evidence that the evaluation did not contain.

Object detection exposes this distinction unusually clearly. The high-end conventional COCO record reviewed here remains near 66 AP after extensive research across model scale, architecture, supervision, data, inference, and search. The relevance of that plateau extends beyond computer vision because

modern detection and language AI now share transformers, sequence modeling, learned representations, gradient optimization, and unified multimodal architectures.

The surface-volume theorem gives the phenomenon a quantitative form. When an open causal state space has effective dimension d and the closed learner is restricted to a codimension-one information boundary, the ratio of available boundary information to required state information tends to $(d-1)/d$. Under the corresponding identification model, Fano's inequality yields the same asymptotic upper bound on exact identification. For the three-dimensional case, the result is two-thirds.

That theorem does not decide the scientific question. Reality does. The scientific hypothesis is that contemporary closed-ended AI serving open-ended objectives approximates this information geometry closely enough to produce a reliability frontier near the predicted value. The object-detection plateau, the ability of closed language benchmarks to exceed the frontier, the collapse of performance as evaluations become more reality-coupled, and the divergence between proxy and external success under recursive optimization are all consistent with that hypothesis.

The analysis concerns the limits of inferring open-world reliability from improvements measured within closed evaluations. Its empirical applicability remains a hypothesis distinct from the mathematical result.

References

[1] T.-Y. Lin et al. Focal Loss for Dense Object Detection. *ICCV, 2017.*

[2] M. Tan, R. Pang, and Q. Le. EfficientDet: Scalable and Efficient Object Detection. *CVPR, 2020.*

[3] Z. Zong, G. Song, and Y. Liu. DETRs with Collaborative Hybrid Assignments Training. *ICCV, 2023.*

[4] Z. Liu et al. Swin Transformer V2: Scaling Up Capacity and Resolution. *CVPR, 2022; preprint first reported in 2021.*

[5] Scylla AI. Scylla AI Leads COCO Benchmark: Best-in-Class Object-Detection Performance and Efficiency. *Developer report, 2025.*

[6] SOTA2 Research. Object Detection on COCO (test-dev). *Secondary index, retrieved 4 November 2026.*

[7] O. Siméoni et al. DINOv3. *arXiv:2508.10104, 2025.*

[8] U. Nisa, M. S. Mohd Pozi, and M. A. Saip. A decade of research in small object detection: a comprehensive bibliometric analysis. *International Journal of Data Science and Analytics, 2025.*

[9] A. H. Roslin and N. Muhammad. Global impact and thematic evolution of object detection in the deep learning era. *Discover Artificial Intelligence 5, 262, 2025.*

[10] arXiv. Advanced-search title-query audit: "object detection," 2023–2026. *Matched January 1–November 3 windows; exact yearly queries reproduced in Appendix A; totals retrieved 4 November 2026.*

[11] W. Wang et al. InternImage: Exploring Large-Scale Vision Foundation Models With Deformable Convolutions. *CVPR, 2023.*

[12] H. Zhang et al. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *ICLR, 2023.*

[13] Y. Li, H. Mao, R. Girshick, and K. He. Exploring Plain Vision Transformer Backbones for Object Detection. *ECCV, 2022.*

[14] Y. Zhao et al. DETRs Beat YOLOs on Real-time Object Detection. *CVPR, 2024.*

[15] A. Wang et al. YOLOv10: Real-Time End-to-End Object Detection. *NeurIPS, 2024.*

[16] J. Huang et al. Revisiting Real-Time Detection Transformer with Efficient Encoder Design. *CVPR Findings, 2026.*

[17] Y. Peng et al. D-FINE: Redefine Regression Task of DETRs as Fine-grained Distribution Refinement. *ICLR, 2025.*

[18] S. Huang et al. DEIM: DETR with Improved Matching for Fast Convergence. *CVPR, 2025.*

[19] S. Chen, P. Sun, Y. Song, and P. Luo. DiffusionDet: Diffusion Model for Object Detection. *ICCV, 2023.*

[20] E. Baty, C. P. Bridges, and S. Hadfield. FlowDet: Unifying Object Detection and Generative Transport Flows. *CVPR Workshops, 2026.*

[21] C. Feng et al. InstaGen: Enhancing Object Detection by Training on Synthetic Dataset. *CVPR, 2024.*

[22] H. Fang et al. Data Augmentation for Object Detection via Controllable Diffusion Models. *WACV, 2024.*

[23] S. Liu et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *ECCV, 2024.*

[24] Q. Chen et al. CP-DETR: Concept Prompt Guide DETR Toward Stronger Universal Object Detection. *AAAI, 2025.*

[25] I. Robinson et al. RF-DETR: Neural Architecture Search for Real-Time Detection Transformers. *ICLR, 2026.*

[26] A. Vaswani et al. Attention Is All You Need. *NeurIPS, 2017.*

[27] A. Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR, 2021.*

[28] N. Carion et al. End-to-End Object Detection with Transformers. *ECCV, 2020.*

[29] T. Chen et al. Pix2Seq: A Language Modeling Framework for Object Detection. *ICLR, 2022.*

[30] B. Xiao et al. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. *CVPR, 2024.*

[31] L. Beyer et al. PaliGemma: A Versatile 3B VLM for Transfer. *arXiv:2407.07726, 2024.*

[32] J. Duchi. Lecture Notes on Statistics and Information Theory. *Stanford University, 2025; covering/metric entropy and Fano's inequality (including Proposition 2.3.3).*

[33] D. Bolya, S. Foley, J. Hays, and J. Hoffman. TIDE: A General Toolbox for Identifying Object Detection Errors. *ECCV, 2020.*

[34] X. Mao et al. COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. *ICCV, 2023.*

[35] S. Penquitt et al. Object Detection Benchmarks are Incomplete: The Role of Label Errors and Annotation Uncertainty. *arXiv:2609.21822, 2026.*

[36] OpenAI. Learning to reason with LLMs. *12 September 2024.*

[37] M. Yuan et al. OSWorld2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks. *arXiv:2606.29537, 2026.*

[38] Z. Wang et al. OSWorld-Pro: Process-based Evaluation for Computer Use Agents. *arXiv:2609.24890, 2026.*

[39] B. Zhao, D. Srikanth, Y. Wu, and Z. Jiang. SpecBench: Measuring Reward Hacking in Long-Horizon Coding Agents. *arXiv:2605.21384, 2026.*

[40] Anonymous authors. Reward Hacking in Self-Improving Code Agents. *ICLR 2026 review manuscript.*

A Object-detection evidence and metric control

The empirical series uses COCO test-dev box AP where available. COCO AP averages precision-recall performance across intersection-over-union thresholds from 0.50 through 0.95. Values from different datasets, splits, prompting protocols, or metrics are recorded separately rather than merged into one leaderboard.

Source	Measurement	Role in the analysis
RetinaNet [1]	39.1 test-dev AP	2017 historical milestone.
EfficientDet-D7	52.2 test-dev AP	2020 historical milestone.
Swin V2 [4]	63.1 test-dev AP	2021/22 large-model milestone.
Co-DETR [3]	66.0 test-dev AP	Published high-end conventional anchor.
ScyllaNet [5]	66.12 reported COCO submission	2025 developer-reported result.
SOTA2 [6]	Co-DETR 66.0 listed highest	Dated persistence observation, November 2026.
DINOv3 [7]	66.1 COCO val with TTA	Same-split saturation comparison, not test-dev.
Co-DETR on LVIS [3]	67.9 LVIS val AP	Different dataset; demonstrates metric/dataset dependence.
CP-DETR [24]	68.4 COCO val AP	Interactive prompting; additional information at inference.

The proximity of the conventional COCO frontier to two-thirds is an empirical observation. Eq. (9) bounds exact state-identification probability, not COCO AP.

Reproducible arXiv activity queries

The activity counts in Section 2 use the exact phrase `"object detection"` in the title, all subject classifications, cross-lists included, and original submission dates. Each window is January 1–November 3. The observed totals on 4 November 2026 were:

Window	Records	Exact query
2023	655	arXiv search
2024	655	arXiv search
2025	626	arXiv search
2026	582	arXiv search

B General dimensional form and metric entropy

For a d-dimensional ball of radius R, a regular ε-cover requires on the order of $(R/\varepsilon)^d$ cells. Its $(d-1)$-dimensional boundary requires on the order of $(R/\varepsilon)^{d-1}$. The logarithms of these covering numbers are metric entropies. Constants depending on the precise geometry disappear in the ratio as R/ε grows.

The argument therefore extends immediately beyond spheres. Any pair of regular metric spaces whose covering numbers scale as

$$N_{\text{open}}(\varepsilon) = \Theta(\varepsilon^{-d}), \qquad N_{\text{closed}}(\varepsilon) = \Theta(\varepsilon^{-q})$$

has an information ratio approaching q/d. The surface-volume model is the codimension-one case $q = d - 1$. The proposed two-thirds frontier is the special case $d = 3$, $q = 2$.

C Shared computation and task-specific information

System	Shared machinery	Information difference
Transformer language model	Attention, token sequence modeling, gradient training [26].	Symbolic input often supplies entities and relations directly.
Vision Transformer	Transformer processing over image patches [27].	Must infer relevant visual structure from pixels.
DETR	Transformer encoder-decoder for class and box prediction [28].	Set matching and localization supervision.
Pix2Seq	Autoregressive token prediction for object classes and coordinates [29].	Detection is represented as conditional language modeling over visual evidence.
Florence-2	Unified sequence-to-sequence model for captioning, detection, grounding, segmentation [30].	Same parameters face tasks with different information requirements.
PaliGemma	Vision Transformer plus Gemma language decoder [31].	Image and text enter one generative model; boxes can be emitted as text-like coordinates.

D Counterexample classification

The counterexample review classified reported results above two-thirds along four axes:

1. whether the task family was fixed before optimization;
2. whether success was determined by a finite known or repeatedly queried verifier;
3. whether the environment can introduce consequential states independently of the evaluator;
4. whether the reported metric measures end-to-end completion or partial/component success.

A benchmark may use realistic artifacts and still be closed. Conversely, a finite experiment can contain stronger reality coupling when the tested environment produces hidden state or dynamic consequences not represented in the optimization loop. This classification concerns closure, not semantic labels such as “realistic” or “artificial.”

The strongest challenge to the theory is a system whose optimization remains closed while prospective, independently generated, end-to-end objectives remain above the predicted frontier across repeated environments. Such a result would challenge the assumed effective dimension, codimension, or information-capacity model.