PREFACE TO THE EDITION

The continuous advancement of Artificial Intelligence has significantly expanded the scope of contemporary research, particularly in the areas of machine learning, intelligent systems, uncertainty estimation, and human–AI interaction. As intelligent technologies increasingly influence critical decision-making, multimodal communication, and autonomous operations, the need for reliable, interpretable, and responsible AI research becomes increasingly important. It is with great pride and academic enthusiasm that we present the latest issue of the **Eduschool Journal of Artificial Intelligence Research (EJAIR)**, featuring scholarly contributions that explore important challenges and emerging directions in modern Artificial Intelligence.

This issue reflects the interdisciplinary nature of Artificial Intelligence by bringing together research on adversarial robustness, conformal prediction, multimodal representation learning, embodied AI, and scalable oversight. The articles collectively examine the reliability, interpretability, adaptability, and responsible development of intelligent systems across diverse computational and real-world applications.

A significant focus of this issue is the development of robust and reliable machine learning systems under changing data conditions. The study on adversarial robustness under distribution shift presents a unified framework connecting natural distribution changes, out-of-distribution detection, and adversarial attacks, while proposing an evaluation protocol to examine classifiers and their detection mechanisms as an integrated system. Equally important is the exploration of uncertainty estimation in high-stakes decision-making. The contribution on conformal prediction examines finite-sample coverage guarantees, calibration procedures, and practical limitations under distribution shifts and subgroup variations, highlighting the importance of interpretable prediction sets in medical, financial, and safety-critical applications.

The issue further explores the advancement of multimodal representation learning through modality-agnostic embeddings that enable cross-modal retrieval across images, text, audio, video, and sensor data. The research examines alignment strategies, representation geometry, retrieval challenges, and scalable search mechanisms, offering valuable insights into the development of universal representations across different modalities. Another important theme addressed in this volume is grounded language generation in embodied AI systems. The article investigates how sensory perception, action-conditioned meaning, and situated language contribute to meaningful interaction with the physical world, while examining existing architectures, available datasets, grounding failures, and emerging research challenges.

Collectively, the contributions in this issue highlight the growing importance of reliability, interpretability, multimodal understanding, and scalable supervision in advancing Artificial Intelligence. The studies presented here provide valuable analytical perspectives and proposed frameworks that encourage further investigation into the technical, practical, and ethical dimensions of intelligent systems.

We extend our sincere gratitude to all authors, reviewers, editorial board members, and contributors whose dedication and scholarly commitment have made this issue possible, strengthening EJAIR as a platform for meaningful academic exchange and innovative research in Artificial Intelligence.

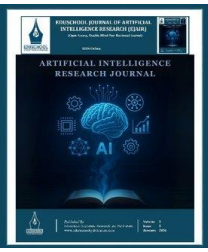
We hope that this issue will inspire researchers, academicians, practitioners, and students to explore emerging research directions and contribute to the development of reliable, interpretable, and responsible Artificial Intelligence systems.

Dr. Juby George
Chief editor

CONTENTS

SL. NO.	TITLE	AUTHOR	PAGE NO.
1	Adversarial Robustness Under Distribution Shift: A Unified View of Detection and Defense	Tintu George	1-8
2	Conformal Prediction for Interpretable Uncertainty in High-Stakes Decision Making	T Ramaprabha	9-16
3	Cross-Modal Retrieval with Modality-Agnostic Embeddings Towards Universal Representation Learning	Rejina P V	17-23
4	Grounded Language Generation from Raw Sensory Input in Embodied AI Systems	Manasy Jayasurya	24-30
5	Scalable Oversight for Advanced AI: Recursive Reward Modeling and Debate	Raji N	31-38

EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Adversarial Robustness Under Distribution Shift: A Unified View of Detection and Defense

Tintu George

Assistant Professor, Department of BCA AI, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India

Article information

Received: 24th July 2026
Received in revised form: 27th August 2026
Accepted: 30th September 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0008

Abstract

Adversarial robustness and out-of-distribution detection have developed as separate research programmes, with separate benchmarks and metrics, although both concern how a classifier behaves on inputs that its training data do not represent. This article reviews them together with work on natural distribution shift and argues that all three study points in one space of test conditions. The space has two axes. The first separates covariate shift, where the label set is unchanged and the correct response is the correct label, from semantic shift, where the input belongs to no training class and the correct response is abstention. The second orders conditions by how strongly the shifted inputs are selected against the model, from random corruption and dataset re-collection, through natural images filtered for model failure, to white-box perturbations chosen by an adaptive adversary. Attack and evaluation methods, natural shift benchmarks, post hoc and training-based detectors, the robustness-accuracy trade-off and certified defenses are reviewed and placed in this space. A relation from the Gaussian noise analysis of adversarial examples is used to connect the random and adversarial ends of the selection axis. An evaluation protocol is then proposed that measures a classifier and its detector as one system in every cell, including attacks written against the detector. The framework and protocol are proposed and have not yet been tested.

Keywords:- Adversarial Robustness, Distribution Shift, Out-of-Distribution Detection, Threat Model, Randomized Smoothing, Evaluation Protocol

.

I. INTRODUCTION

A deep network trained on a fixed distribution is evaluated, in most published work, on a held-out sample of that same distribution. Deployment rarely respects this assumption. Images arrive blurred, compressed or taken under unfamiliar lighting, objects appear in renditions never seen during training, and some inputs belong to no class the model was asked to recognise. A further kind of input is constructed on purpose, since small perturbations found by an optimiser can change the prediction of an accurate image classifier while leaving the image visually unchanged [1], [2]. Each situation has produced its own literature.

The three literatures use different measures. Adversarial robustness is measured as accuracy under a worst-case perturbation inside an $\ell\infty$ or $\ell2$ ball. Most defenses are built on the min-max formulation of adversarial training [3]. Robustness to natural shift is measured on corrupted or re-collected test sets such as ImageNet-C [4] and ImageNetV2 [5]. Out-of-distribution (OOD) detection is measured by how well a confidence score separates in-distribution test images from images of unseen classes, starting from the maximum softmax probability baseline

[6]. The three communities report different metrics, and a model presented as strong under one of them is seldom evaluated under the other two.

There are good reasons to treat them together. One early detector, based on class-conditional Gaussian models of network features, was shown to flag both unseen-class inputs and adversarial examples with the same score [7]. An analysis of Gaussian noise has tied the error rate of a classifier under random corruption to the distance from a clean input to its nearest misclassified point [8]. A detector is itself a function of the input. As a result, an attacker who knows it can perturb an OOD input until the detector accepts it. A deployed system therefore faces natural shift, semantic novelty and deliberate manipulation at the same time, and an evaluation that measures them one at a time cannot see the interactions between them.

This article reviews the three literatures and proposes a unified view of them, beginning in Section II with adversarial attacks and with the evaluation failures that shaped current practice in that field. Section III covers natural distribution shift, and Section IV covers OOD detection and its exposure to attack. Section V proposes a threat model with two axes, the type of shift and the strength of selection against the model. Section VI discusses defenses, the robustness-accuracy trade-off and certification. Section VII proposes an evaluation protocol and states the limitations of the proposal.

II. ADVERSARIAL EXAMPLES AND THE EVALUATION PROBLEM

Szegedy et al. reported in 2014 that imperceptible perturbations could cause misclassification [1]. They also found that the same perturbed images often fooled networks trained separately [1]. Goodfellow et al. attributed the effect to the nearly linear behaviour of networks in high dimension [2], and their fast gradient sign method (FGSM) moves the input one step of size ε along the sign of the loss gradient. Madry et al. recast attack and defense as one saddle point problem [3], in which the objective is the minimum over parameters θ of the expected maximum over $\|\delta\| \le \varepsilon$ of the loss $\mathcal{L}(f_\theta(x+\delta), y)$. Projected gradient descent (PGD) solves the inner maximisation. Carlini and Wagner wrote attacks as optimisation problems under $\ell 0$, $\ell 2$ and $\ell\infty$ distances and used them to defeat defensive distillation, which had resisted the earlier attacks [9].

Much of the later history concerns evaluation. Athalye et al. examined the non-certified white-box defenses accepted at ICLR 2018 and found that seven of the nine relied on obfuscated gradients, a form of gradient masking that defeats gradient-based attacks without removing the adversarial examples [10]. With attacks adapted to each mechanism they circumvented six of the seven completely and one partially. Tramer et al. later broke thirteen published defenses, each with an attack written for that defense, and set out the methodology an adaptive attack requires, arguing that each attack has to be tailored to the defense rather than run as a fixed procedure [11]. A defense therefore has to be evaluated by an attacker who knows it. An added OOD detector enlarges the surface that such an attacker can target.

Standardised attacks removed the worst cases of weak evaluation. AutoAttack has four components [12]. Two are parameter-free variants of PGD with an automatic step size schedule, and the other two are the minimum-norm FAB attack and the query-based Square attack. Applied to more than fifty published models, it gave a lower robust accuracy than the original papers in all but one case, often by more than ten percentage points [12]. RobustBench maintains leaderboards built on AutoAttack for $\ell\infty$ and $\ell 2$ threat models and for common corruptions [13]. They cover CIFAR-10, CIFAR-100 and ImageNet. Entries are restricted to models for which the attack is known to be reliable, which excludes randomised defenses and defenses that run an optimisation loop at test time. Almost all threat models in this literature are ℓp balls around a clean input, chosen for tractability. Real adversaries are not confined to such balls.

III. NATURAL DISTRIBUTION SHIFT

Most natural shift benchmarks for vision are built on ImageNet [14] and keep its label set or a subset of it, so that the same classifier can be tested without retraining. ImageNet-C applies fifteen corruption types, grouped into noise, blur, weather and digital artefacts, at five severity levels each [4]. Each validation image therefore has 75 corrupted versions. The benchmark reports corruption error normalised by that of AlexNet. ImageNetV2 repeats the original collection procedure on new images and finds accuracy drops of 11% to 14% for ImageNet models, which the authors attribute to the new images being slightly harder rather than to adaptive overfitting on the original test set [5].

Other benchmarks change the kind of image. ImageNet-R contains 30,000 renditions of 200 ImageNet classes, including paintings, cartoons, sculptures and embroidery [15]. ImageNet-A contains real, unmodified photographs gathered by adversarial filtration, which keeps the web images that a fixed classifier gets wrong, and its companion ImageNet-O applies the same filtering to images of classes outside ImageNet to build an OOD detection set [16]. On ImageNet-A a DenseNet-121 reaches about 2% accuracy, and its OOD detection on

ImageNet-O is near chance. WILDS collects ten datasets with shifts that occur in practice, among them hospitals, camera trap locations and satellite regions, and defines its splits by those domains [17].

Two empirical regularities organise this literature. Taori et al. evaluated 204 ImageNet models under 213 test conditions [18]. They found little or no transfer of robustness from synthetic shifts to natural ones, with training on larger and more diverse data as the main exception [18]. Miller et al. found that, across many models and shifts, accuracy on the shifted distribution is close to a linear function of in-distribution accuracy after a probit transform of both [19]. A method improves robustness only if it lifts shifted accuracy above the line predicted from its in-distribution accuracy, a quantity Taori et al. call effective robustness [18]. Hendrycks et al. reached a less negative conclusion on their own datasets [15]. They found that larger models and some artificial augmentations improve robustness to real shifts but that no evaluated method helps consistently [15]. The usual explanation for poor transfer is shortcut learning, in which a network adopts decision rules, such as reliance on texture or background, that suffice on the benchmark and fail outside it [20].

IV. OUT-OF-DISTRIBUTION DETECTION AND ITS EXPOSURE TO ATTACK

OOD detection asks for a score s(x) that is high for in-distribution inputs and low for inputs whose class lies outside the training label set. The maximum softmax probability (MSP) of a trained classifier gives the standard baseline [6]. It needs no retraining. ODIN sharpens the separation with temperature scaling of the softmax and a small input perturbation that increases the maximum softmax score, a step with the same form as FGSM taken in the opposite direction [21]. The Mahalanobis detector fits class-conditional Gaussians with a shared covariance to intermediate features and scores an input by its distance to the closest class mean [7]. The energy score replaces the softmax maximum with :

$$E(x) = -T \log \sum_i \exp\left(\frac{f_i(x)}{T}\right) \quad (1)$$

computed from the logits $f_i(x)$ at temperature T [22].

Later post hoc methods clip penultimate activations before computing the score [23], or use the distance from an input to its k-th nearest training feature in a normalised embedding [24]. Outlier exposure takes a different route and trains the classifier to give uniform predictions on an auxiliary dataset of outliers [25]. Table 1 lists these families.

Table 1. Families of out-of-distribution detector

Family	Score computed from	Retraining needed	Auxiliary outliers	Representative work
Softmax confidence	Maximum softmax probability	No	No	[6]
Temperature and input perturbation	Scaled softmax after one gradient step	No	For tuning	[21]
Feature density	Mahalanobis distance to class means	No	For tuning	[7]
Energy	Log-sum-exp of the logits	Optional	Optional	[22]
Activation shaping	Clipped penultimate activations	No	No	[23]
Nearest neighbour	k-th neighbour distance in feature space	No	No	[24]
Outlier exposure	Confidence after training on outliers	Yes	Yes	[25]
Adversarial confidence training	Confidence on perturbed outliers	Yes	Synthetic or real	[26], [27]

OpenOOD gathers detectors and datasets under one protocol [28]. It separates near-OOD sets, whose classes resemble the training classes, from far-OOD sets [28]. Yang et al. place OOD detection beside anomaly detection, novelty detection, open set recognition and outlier detection in a single framework of generalised OOD detection [29]. Large pretrained transformers raise performance on hard near-OOD pairs such as CIFAR-100 against CIFAR-10 [30]. Apart from the last row, none of the scores in Table 1 was designed with an attacker in mind, and each is computed from a network that an attacker can differentiate through.

In 2019 Hein et al. proved that ReLU networks produce arbitrarily high-confidence predictions on inputs far from the training data, because scaling an input by a large factor drives the logits along a fixed linear direction [26]. Carlini and Wagner bypassed ten detectors of adversarial examples by adding each detector to the attack objective [31]. The same construction applies to OOD detectors. A perturbation that raises s(x) on an OOD input turns it into an accepted input, and a perturbation that lowers s(x) on an in-distribution input causes a false alarm. Provable guarantees exist for restricted settings, such as low confidence far from the training data under a density model [32] and low confidence within an $\ell\infty$ ball around OOD inputs using interval bound propagation [27]. These guarantees hold at small radii and have been shown on small datasets such as CIFAR-10.

V. A UNIFIED THREAT MODEL

The framework proposed here describes any test condition by two coordinates. The first, following the distinction drawn in [29], is the type of shift. Covariate shift changes the input distribution while keeping the label set Y fixed, so the correct response is the correct label. Semantic shift supplies inputs whose true class lies outside Y, so the correct response is abstention. The second coordinate is the strength of selection against the model, which runs from random sampling of shifted inputs, through filtering of natural inputs by model failure, to white-box optimisation by an adversary who knows both the classifier and the detector. Fig. 1 shows the resulting grid with existing benchmarks placed in its cells.

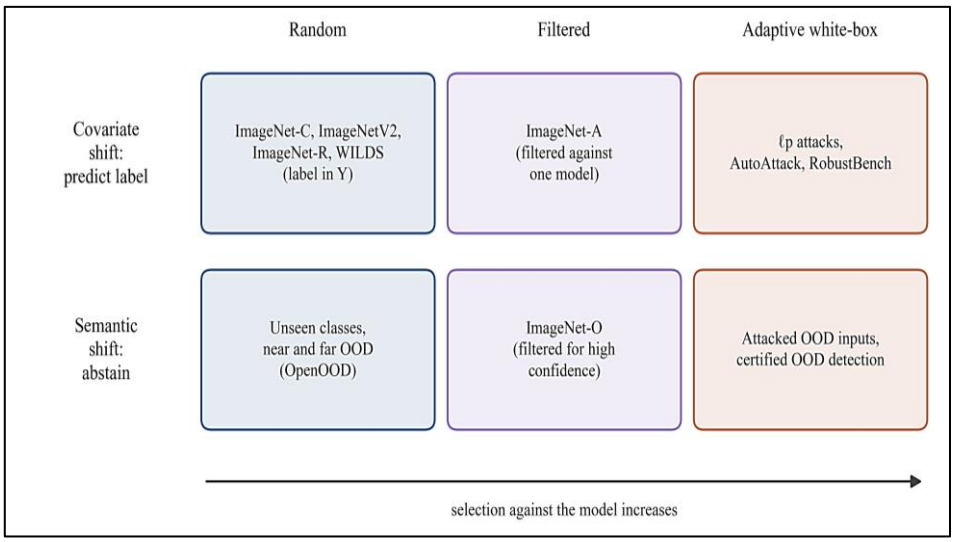


Fig. 1: Schematic of the proposed two-axis threat model. Rows give the type of shift and columns give the strength of selection against the model, with representative benchmarks in each cell: corruptions, renditions and re-collected sets [4], [15], [5], [17], filtered sets [16], norm-bounded attacks [12], [13], unseen-class benchmarks [28] and attacked OOD inputs [27].

Formally, a test condition is a triple (Q, Δ, A), where Q is a source distribution, Δ(x) is the set of admissible modifications of an input x, and A is the rule that selects one element of Δ(x). The system under test is a pair (f, s) with a threshold τ. It returns f(x) when $s(x) \geq \tau$ and abstains otherwise. Its risk is the expected loss of the selected input, with x drawn from Q, where the loss has separate costs for a wrong label on an accepted covariate input, an accepted semantic input and a rejected covariate input, including a perturbed one.

Standard adversarial robustness is the case Q = P, the training distribution, with A the worst case over an ℓp ball. Natural shift robustness takes A as the identity and Q as a corrupted or re-collected distribution. OOD detection also keeps A as the identity but takes Q supported outside Y. Table 2 records these cases and the intermediate ones.

Table 2. Placement of test conditions in the proposed two-axis threat model

Condition	Shift type	Selection	Admissible set Δ(x)	Correct response	Examples
In-distribution test	None	Random	The input itself	Label	[14]
Common corruptions	Covariate	Random	15 corruptions, 5 severities	Label	[4]
Re-collection and renditions	Covariate	Random	New sample from Q	Label	[5], [15], [17]
Filtered natural images	Covariate	Against one fixed model	Unmodified natural images	Label	[16]
Norm-bounded attack	Covariate	Adaptive white-box	ℓp ball of radius ε	Label	[3], [12]
Attack on the detector	Covariate	Adaptive white-box	ℓp ball, score as target	Label, no abstention	[31]
Unseen classes	Semantic	Random	New sample outside Y	Abstain	[6], [28]
Filtered OOD images	Semantic	Against one fixed model	Unmodified natural images	Abstain	[16]
Attacked OOD inputs	Semantic	Adaptive white-box	ℓp ball around an OOD input	Abstain	[26], [27]

The two ends of the selection axis are not independent. Gilmer et al. considered isotropic Gaussian noise with standard deviation σ per coordinate and the error rate μ of a classifier on noisy copies of a clean input x [8]. If the set of misclassified points is a half space, the ℓ2 distance from x to its nearest error equals:

$$\varepsilon^* = -\sigma\Phi^{-1}(\mu) \tag{2}$$

where Φ is the standard normal distribution function. Through the Gaussian isoperimetric inequality the half space is the most favourable shape, so error sets of other shapes with the same measure lie, in a median sense, no further away from the noisy copies of x [8]. Fig. 2 plots ε* for three noise levels. At σ = 0.1 an error rate of 1% under noise already places an error within an ℓ2 distance of about 0.23. At μ = 10% the distance falls to about 0.13.

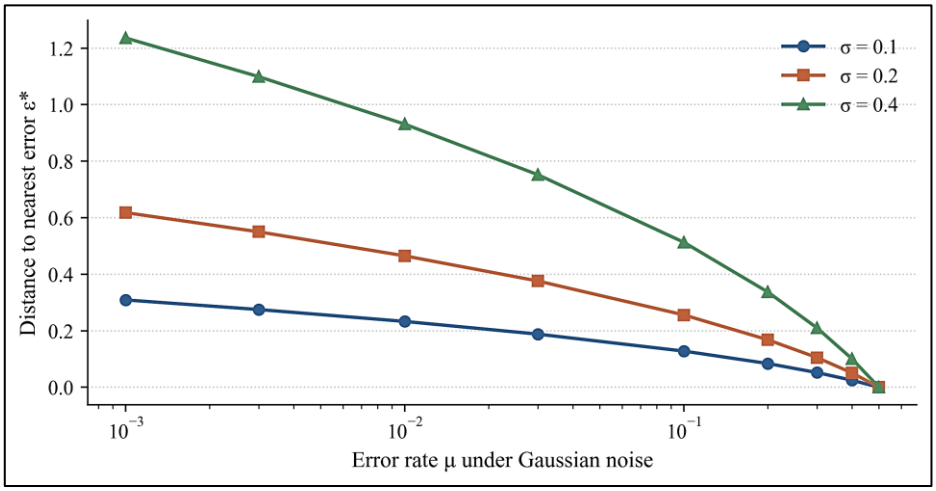


Fig. 2: Distance from a clean input to its nearest error, $\varepsilon^* = -\sigma\Phi^{-1}(\mu)$, as a function of the error rate μ under isotropic Gaussian noise with standard deviation σ, computed from the half space relation of [8]. The horizontal axis is logarithmic.

A model with measurable error under random noise is therefore expected, under the half-space approximation, to have adversarial errors close by, and Gilmer et al. accordingly treat adversarial and corruption robustness as two views of one phenomenon [8]. ImageNet-A and ImageNet-O were built by keeping natural images on which a fixed model failed [16]. The adversary here selects inputs without modifying them and uses no gradients. It therefore sits between random shift and white-box attack. The rows of Table 2 for attacks on the detector and for attacked OOD inputs are the cells that current practice leaves sparsely populated.

VI. DEFENSES, TRADE-OFFS AND CERTIFICATION

Adversarial training with PGD and its variants underlies most of the leading entries on the RobustBench leaderboards [3], [13]. It costs clean accuracy. Tsipras et al. constructed a simple distribution on which any classifier with high standard accuracy must have low robust accuracy [33]. TRADES writes the robust error as

the sum of the natural error and a boundary error, and trains with a cross-entropy term plus a Kullback-Leibler term between the predictions on x and on its worst-case perturbation, weighted by a parameter that sets the balance [34].

In the two-axis view the trade-off has a second dimension. Adversarial training constrains covariate inputs within an ℓp ball. It places no constraint on semantic inputs, and a model trained only to keep its label inside the ball has no reason to lower its confidence on an unseen class. Outlier exposure trains low confidence on semantic inputs without any constraint on perturbations [25]. Hein et al. combined the two by enforcing low confidence on perturbed noise inputs [26], and Bitterwolf et al. made this certifiable within ℓ∞ balls around outliers [27]. No published defense known to the authors covers all six cells of Fig. 1. The cost of defending one cell in terms of another has been reported mainly for single pairs of cells, most often clean accuracy against ℓ∞ robustness.

Certification replaces an empirical attack with a guarantee. Randomized smoothing classifies a point by the most probable output of a base classifier under Gaussian noise with standard deviation σ. Cohen et al. proved that the smoothed classifier is constant within the ℓ2 radius:

$$R = \frac{\sigma}{2}[\Phi^{-1}(p_A) - \Phi^{-1}(p_B)] \quad (3)$$

where p_A is a lower bound on the probability of the top class and p_B an upper bound on that of the runner-up [35]. The bound is tight for Gaussian noise. With $p_B = 1 - p_A$ the radius becomes $\sigma\,\Phi^{-1}(p_A)$. Fig. 3 plots this radius for the three noise levels Cohen et al. used on ImageNet.

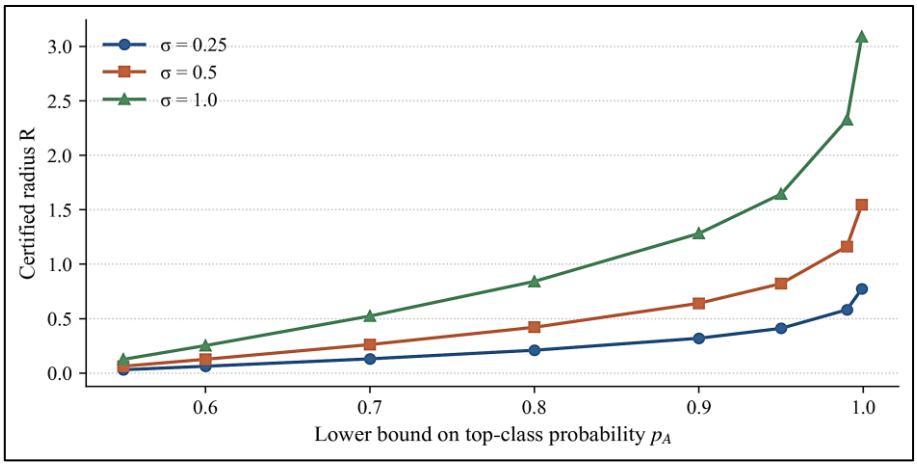


Fig. 3: Certified ℓ2 radius of a Gaussian smoothed classifier, $R = \sigma\Phi^{-1}(p_A)$, for the case $p_B = 1 - p_A$, computed from the bound of [35] for the three noise levels used there on ImageNet.

In Fig. 3 the radius exceeds σ only when p_A passes $\Phi(1)$, about 0.84. It grows slowly after that, reaching 2.33σ at $p_A = 0.99$. A larger σ buys a larger radius but makes the base classifier work on noisier inputs, and on ImageNet Cohen et al. reported a certified top-1 accuracy of 49% at an ℓ2 radius of 0.5 [35]. Smoothing turns accuracy under Gaussian noise, a random covariate condition, into a guarantee against white-box ℓ2 attack. This is a converse of the relation of Fig. 2, applied to the smoothed classifier. Pairing a smoothed classifier with a smoothed detector score has received less study than the smoothed classifier alone. Certified radii of this size are also small compared with the change that a rendition or a weather corruption makes to an image.

VII. PROPOSED EVALUATION PROTOCOL AND LIMITATIONS

The protocol proposed here reports a classifier and its detector together, as one system (f, s, τ), in every cell of Fig. 1. Six measurements are required. Table 2 lists the cells they cover. Clean accuracy on the in-distribution test set fixes the reference point. Accuracy under random covariate shift is reported per corruption and severity and on re-collected sets, together with its distance above or below the linear trend of shifted against clean accuracy, so that gains are not credited to higher clean accuracy alone [18], [19]. Accuracy on filtered sets is reported with the identity of the filtering model. Robust accuracy under AutoAttack is reported at several radii rather than one [12]. Detection of semantic shift is reported separately for near-OOD and far-OOD sets under a shared protocol such as OpenOOD [28]. The sixth measurement attacks the detector, perturbing in-distribution inputs to lower s(x) and OOD inputs to raise it, with the attack written against the whole system, including any input preprocessing such as the ODIN step [21], [11].

A single summary needs a fixed threshold. The protocol sets τ on clean in-distribution validation data at a stated true positive rate, such as 95%. It is then held fixed across all cells. For each cell it then reports the

acceptance rate and the selective risk, which is the error rate among accepted inputs. Where a certificate exists, the certified fraction at each radius is reported beside the empirical figure, following the practice of Cohen et al. for accuracy [35].

Five limitations apply. The framework has not been applied to any model, so the expectation that it changes the ranking of methods relative to single-axis evaluation is untested. The selection axis is ordinal rather than metric, and two filtered datasets built against different models cannot be compared in strength. Semantic shift is defined relative to a label set, while many deployed inputs are partly novel, for example a known object in an unfamiliar context. The binary first axis does not describe such inputs. The attacks in the sixth measurement have to be adapted to each detector, which makes the protocol expensive, and like every empirical robustness figure its results give only an upper bound on true performance [11]. Almost all the evidence reviewed here comes from image classification, and its extension to language models, where shift and attack act on text rather than pixels, is open.

VIII. CONCLUSION

Adversarial robustness, robustness to natural shift and OOD detection can be read as studying the same object, the behaviour of a model away from its training distribution. They differ in the type of shift and in how strongly the shifted inputs are selected against the model. Most defenses cover one or two cells of the grid in Fig. 1, detectors are usually reported without the adaptive attacks that adversarial robustness research learned to require, and, under a half-space approximation, the relation plotted in Fig. 2 ties the random and adversarial ends of the selection axis together.

The proposed protocol measures one system in every cell, with a fixed threshold and with attacks written against the detector. Running it on the public models collected by RobustBench and OpenOOD is the next step [13], [28]. Safety methods that act on model behaviour, such as alignment through self-critique and revision as discussed in this journal [36], assume that a model behaves on deployment inputs as it did during development. The measurements proposed here give one way to check that assumption.

REFERENCES

[1] C. Szegedy et al., “Intriguing properties of neural networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[4] D. Hendrycks and T. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[5] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do ImageNet classifiers generalize to ImageNet?,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 5389–5400.

[6] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[7] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.

[8] J. Gilmer, N. Ford, N. Carlini, and E. Cubuk, “Adversarial examples are a natural consequence of test error in noise,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 2280–2289.

[9] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2017, pp. 39–57.

[10] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 80, 2018, pp. 274–283.

[11] F. Tramer, N. Carlini, W. Brendel, and A. Madry, “On adaptive attacks to adversarial example defenses,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1633–1645.

[12] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 119, 2020, pp. 2206–2216.

[13] F. Croce et al., “RobustBench: A standardized adversarial robustness benchmark,” in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, 2021.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[15] D. Hendrycks et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8320–8329.

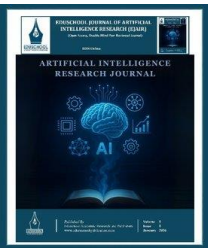
[16] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, “Natural adversarial examples,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15257–15266.

[17] P. W. Koh et al., “WILDS: A benchmark of in-the-wild distribution shifts,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, 2021, pp. 5637–5664.

[18] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 18583–18599.

[19] J. P. Miller et al., "Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, 2021, pp. 7721–7735.
[20] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.
[21] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
[22] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 21464–21475.
[23] Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-distribution detection with rectified activations," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 144–157.
[24] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 162, 2022, pp. 20827–20840.
[25] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
[26] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 41–50.
[27] J. Bitterwolf, A. Meinke, and M. Hein, "Certifiably adversarially robust detection of out-of-distribution data," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 16085–16095.
[28] J. Yang et al., "OpenOOD: Benchmarking generalized out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 32598–32611.
[29] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *Int. J. Comput. Vis.*, vol. 132, no. 12, pp. 5635–5662, 2024.
[30] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 7068–7081.
[31] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. 10th ACM Workshop Artif. Intell. Secur. (AISec)*, 2017, pp. 3–14.
[32] A. Meinke and M. Hein, "Towards neural networks that provably know when they don't know," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
[33] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
[34] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 7472–7482.
[35] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 1310–1320.
[36] W. M. John, "Constitutional AI: Scalable alignment through self-critique and revision," *Eduschool J. Artif. Intell. Res.*, vol. 1, no. 1, pp. 10–19, 2026, doi: 10.63090/EJAIR/3139.7336.0002.

EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Conformal Prediction for Interpretable Uncertainty in High-Stakes Decision Making

T Ramaprabha

Associate Professor, Department of Computer Science, Nehru Arts and Science College, Coimbatore, India

Article information

Received: 29th July 2026
Received in revised form: 31st August 2026
Accepted: 1st October 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0009

Abstract

Deep networks used in medicine, finance and autonomous systems report confidence scores that are often miscalibrated and that carry no guarantee once the data depart from the training distribution. Conformal prediction replaces the single score with a set of candidate outcomes and attaches a finite-sample guarantee to it. Under exchangeability, the set contains the true outcome with probability at least $1 - \alpha$, for any model and any data distribution. This article reviews the method as a tool for high-stakes decisions. It sets out the split conformal procedure and its coverage bounds, computes how the guarantee behaves at the calibration sizes available in clinical and safety-critical settings, and surveys the nonconformity scores used for classification and regression. It then examines three points at which the basic guarantee falls short in practice, namely conditional coverage across patients or subgroups, distribution shift, and losses other than miscoverage, and reviews the extensions that address each. Applications in medical imaging, pathology and language generation are assessed with the evidence on human use of prediction sets. The article closes with a proposed reporting protocol that ties the choice of α, the calibration data and the behaviour of set size to the decision the set is meant to support. The protocol has not been applied to a deployed system and is offered for testing.

Keywords:- Conformal Prediction, Uncertainty Quantification, Prediction Sets, Calibration, Distribution Shift, Risk Control, High-Stakes Decision Making

I. INTRODUCTION

A model that classifies a chest radiograph, flags a fraudulent transaction or proposes a drug dose produces a number alongside its answer, and that number is usually read as the probability that the answer is right. For modern deep networks this reading is unsafe. Guo et al. found that modern deep networks are systematically overconfident [1]. Their benchmarks covered image and text classification. A single temperature parameter fitted after training removed much of the gap on held-out data from the same distribution. Later work found that calibration depends strongly on architecture [2]. Popular recalibration methods are less calibrated than reported, and standard estimators cannot tell by how much [3]. Post-hoc calibration also falls short under dataset shift [4]. Calibration measured on a test set is an average property of the score, and it gives no statement about the error rate to expect on the cases a clinician sees next week.

Conformal prediction takes a different route [5], [6]. Instead of a point prediction with a score, it returns a set of labels or an interval, built so that the set contains the true outcome with a probability fixed in advance by the user. The guarantee holds in finite samples and for any underlying model, provided the calibration data and the new case are exchangeable. Two properties make the method attractive in high-stakes settings. The guarantee

does not require the model to be correct, so a poor model tends to yield larger sets on average rather than false confidence. The size of the set is also a readable statement of uncertainty, because a set with one label, a set with four labels and an empty set each carry a meaning on which a decision maker can act [7].

These properties are widely promoted, while the limits that accompany them receive less attention in applied work. Marginal coverage averaged over a population can coexist with poor coverage for a subgroup [8]. Exchangeability fails under the shifts that deployment brings [9]. The quantity a hospital cares about may also be a false negative rate rather than miscoverage [10]. This article reviews the method with these limits in view. Section II covers confidence scores and calibration. Section III states the split conformal procedure and its finite-sample behaviour. Section IV covers conditional coverage, shift and risk control. Section V reviews applications and human use of prediction sets, Section VI proposes a reporting protocol, and Section VII concludes.

II. CONFIDENCE SCORES WITHOUT GUARANTEES

A classifier with softmax output is calibrated when, among all inputs that receive confidence p for the predicted class, a fraction p is classified correctly. The property is defined over the input distribution. It is usually estimated by binning predictions and comparing average confidence with accuracy in each bin, which gives the expected calibration error used by Guo et al. [1]. Temperature scaling divides the logits by a scalar T fitted on a validation set. The ranking of classes is unchanged. A value of T above 1 softens the probabilities and lowers overconfidence. The method is simple and works well on data from the training distribution. It has become the default post-hoc correction.

Its value for decisions is limited in three ways, beginning with estimation. Kumar et al. showed that Platt scaling and temperature scaling are less calibrated than binned estimates suggest, and that the usual estimators cannot measure the remaining miscalibration, so a reported error near zero offers little assurance [3]. Dataset shift is a separate problem. Ovadia et al. compared uncertainty methods in a large benchmark under dataset shift [4]. In their results, post-hoc calibration on held-out data did not survive the shift. Methods that average over several models, such as ensembles of independently trained networks [11], gave stronger results than single networks with temperature scaling or Monte Carlo dropout [12]. Calibration also concerns averages over inputs that share a score, and it does not translate into a bound on the error rate of a specific decision rule, such as acting on the top label whenever confidence exceeds 0.9.

Bayesian approximations and ensembles give richer estimates of uncertainty. Their credible regions, however, are valid only if the posterior approximation is accurate, which is hard to check for networks with millions of parameters. High-stakes use needs a weaker and more practical statement, namely that a stated procedure misses the true outcome in at most a fixed fraction of cases, together with an output whose size carries the uncertainty. Rudin has argued that high-stakes decisions call for models whose reasoning can be inspected [13]. Kompa et al. argued that clinical models should quantify and communicate uncertainty and should be able to abstain on uncertain cases [14]. Conformal prediction meets the coverage requirement directly and supplies a set as its output. The set does not, however, explain the reasoning of the underlying model.

III. SPLIT CONFORMAL PREDICTION IN FINITE SAMPLES

A. The procedure and its guarantee

The split, or inductive, form of conformal prediction separates model fitting from calibration [15], [16]. A model is trained on one portion of the data. A nonconformity score s(x, y) is defined that grows as a candidate label y fits the input x less well. A common choice for classification is $s(x, y) = 1 - f_y(x)$, one minus the softmax probability of label y [17]. The score is computed on n held-out calibration pairs, giving $s_1, \ldots, s_n$. The threshold q is the $\lceil (n + 1)(1 - \alpha) \rceil$-th smallest of these scores, and the prediction set for a new input x is $C(x) = \{y: s(x, y) \leq q\}$. Any trained classifier can supply the score, and the model itself is not retrained.

If the calibration pairs and the test pair are exchangeable, the rank of the test score among all n + 1 scores is uniformly distributed. The set then satisfies:

$$1 - \alpha \leq P(Y \in C(X)) \leq 1 - \alpha + \frac{1}{n+1} \quad (1)$$

where the upper bound requires continuous scores without ties [7], [5]. The guarantee is marginal. It averages over the draw of the calibration set and the test point. It makes no claim about any particular input. Fig. 1 plots the empirical quantile level $\frac{\lceil (n + 1)(1 - \alpha) \rceil}{n}$ that the threshold must reach, for $\alpha = 0.05$, 0.1 and 0.2. The level exceeds $1 - \alpha$ by a margin that shrinks as n grows. At $n = 20$ and $\alpha = 0.1$ it equals 0.95. The threshold is finite only when $n \geq \frac{1}{\alpha} - 1$, so at least 19 calibration points are needed for $\alpha = 0.05$ and at least 9 for $\alpha = 0.1$.

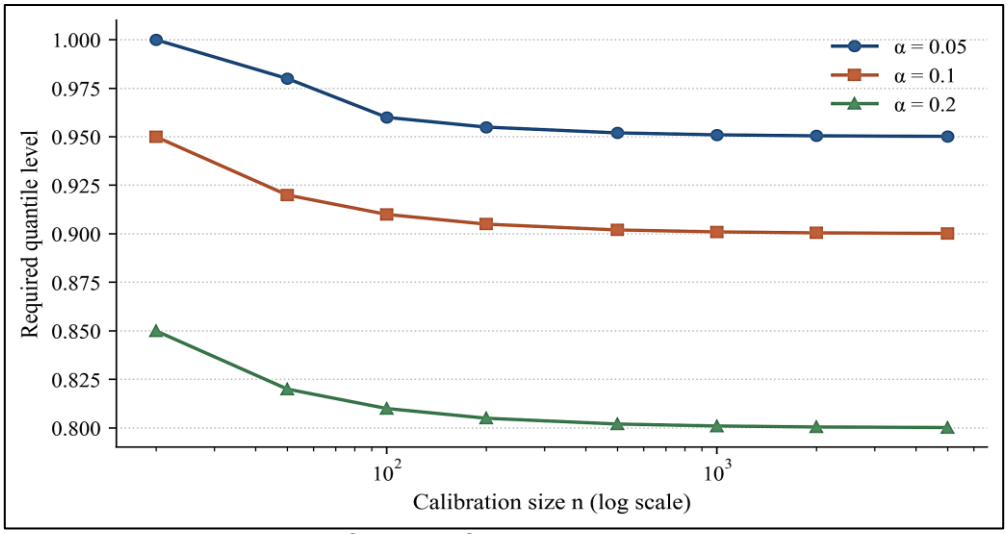


Fig. 1: Finite-sample quantile level $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ at which split conformal prediction sets its threshold, computed from the formula in Section III for calibration sizes n from 20 to 5000 and three miscoverage levels α. The level approaches the nominal $1 - \alpha$ as n grows.

B. Coverage given a fixed calibration set

A second distinction matters more for deployment. A hospital calibrates once and then applies the same threshold for months. The relevant quantity is then the coverage conditional on the particular calibration set that was drawn. For i.i.d. data and continuous scores, this conditional coverage follows a Beta distribution with parameters $n + 1 - l$ and l, where $l = \lfloor (n + 1)\alpha \rfloor$ [18], [7]. Its mean is close to $1 - \alpha$, but its spread is wide at small n. Fig. 2 plots the 5th, 50th and 95th percentiles of this distribution for $\alpha = 0.1$. The median stays close to 0.90 throughout. With $n = 100$, the 5th percentile is 0.848 and the probability that the realised coverage falls below 0.88 is 0.23, whereas with $n = 1000$ the 5th percentile rises to 0.884 and that probability falls to 0.02. Calibration sets of a few hundred cases are not unusual when labels need expert review, and at that size the coverage actually delivered can differ from the nominal level by several percentage points in either direction.

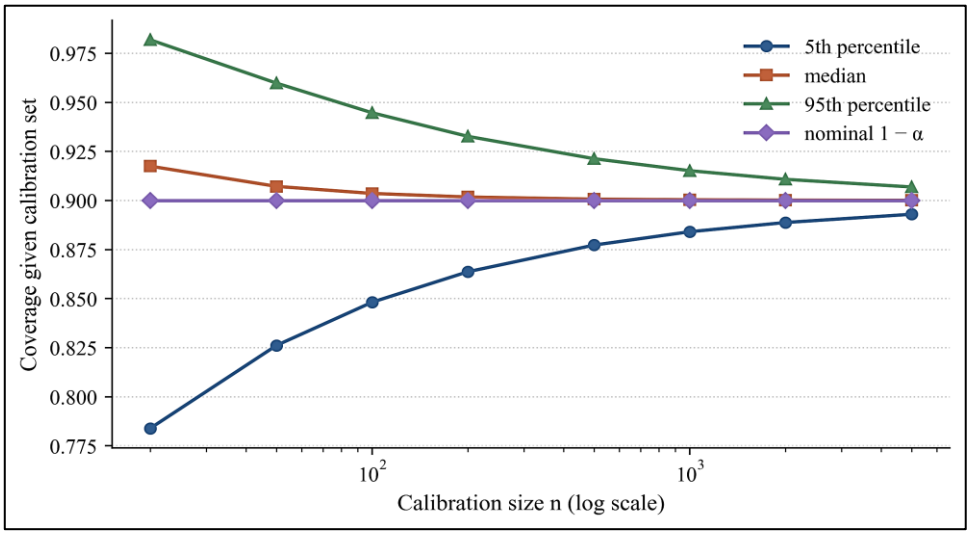


Fig. 2: Coverage conditional on the calibration set for split conformal prediction at $\alpha = 0.1$, computed from the Beta($n + 1 - l$, l) law with $l = \lfloor (n + 1)\alpha \rfloor$ stated in Section III [18], [7]. Lines show the 5th, 50th and 95th percentiles for n from 20 to 5000, with the nominal level for reference.

C. The choice of score

Any score yields valid sets. The score decides only how useful the sets are. Table 1 lists six of them. The threshold score of Sadinle et al. gives the smallest expected set size among procedures with the same coverage when the estimated class probabilities are correct [17]. In practice it tends to under-cover hard inputs and over-cover easy ones [7], [19]. Adaptive prediction sets (APS) add sorted class probabilities until the true label is reached, which gives larger sets on ambiguous inputs and better approximate conditional coverage [19]. The regularised variant RAPS penalises the long tail of unlikely classes, and on ImageNet its authors report sets often

5 to 10 times smaller than a Platt scaling baseline while achieving the target coverage [20]. For regression, conformalized quantile regression (CQR) calibrates the band between two fitted quantiles, so that interval width varies with the input [21], whereas the absolute residual score gives intervals of constant width [16].

Table 1. Nonconformity scores used in conformal prediction

Score	Task	Definition	Behaviour of the sets	Source
Threshold	Classification	One minus the softmax probability of the label	Smallest average size, uneven coverage across difficulty	[17]
APS	Classification	Cumulative sorted probability up to the label, randomised	Larger sets on ambiguous inputs	[19]
RAPS	Classification	APS with a penalty on labels beyond rank k	Smaller and more stable sets	[20]
Mondrian	Classification	Any score, calibrated within each class or group	Coverage within every category, fewer data per category	[5], [22]
Absolute residual	Regression	$\|y - \mu(x)\|$ for a fitted mean μ	Intervals of constant width	[16]
CQR	Regression	Distance outside a fitted quantile band	Width adapts to the input	[21]

IV. CONDITIONAL COVERAGE, SHIFT AND RISK CONTROL

A. Conditional and group coverage

Marginal coverage permits a procedure to cover 99% of easy cases and 70% of hard ones while averaging 90%, when about two thirds of cases are easy. In clinical use the hard cases are often those with the greatest consequences. Exact coverage conditional on every input value cannot be obtained without assumptions on the distribution [18]. Vovk analysed which forms of conditional validity inductive conformal predictors can and cannot achieve [18]. Foygel Barber et al. state the result that exact distribution-free conditional coverage forces intervals of infinite expected length at almost every point that is not an atom of the feature distribution [8]. They then examined weaker targets that remain achievable.

Coverage within groups is attainable when the groups are defined in advance. The Mondrian construction calibrates separately within each category, such as each class label or each patient subgroup, and gives the guarantee within every category at the cost of smaller calibration sets [5]. Romano et al. used this idea to obtain equalized coverage across protected groups [23]. Lu et al. interviewed radiologists about possible uses of conformal predictors, evaluated several conformal methods for skin lesion classification and adapted them to give equal coverage across groups of patient skin tone [24]. When the number of classes is large, calibration within each class runs short of data [22]. Ding et al. proposed clustering classes with similar score distributions so that each cluster has enough calibration examples [22]. The Beta law of Fig. 2 applies within each group, so a subgroup with 100 calibration cases inherits the wide spread computed for n = 100.

B. Distribution shift

Exchangeability fails when the deployed population differs from the calibration population, which is the normal condition once a model leaves the site where it was built. Tibshirani et al. showed that under covariate shift, where the distribution of inputs changes but the conditional distribution of outcomes does not, coverage is restored by weighting calibration scores with the likelihood ratio between the test and calibration input densities [9]. The ratio must be known or estimated, and errors in the estimate pass directly into the coverage. For data that arrive in sequence, adaptive conformal inference updates the working level after each case [25]. The update is:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t) \tag{2}$$

Here err_t is 1 when the set at time t missed the outcome and 0 otherwise. It needs no model of the shift. The long-run miscoverage frequency then converges to α for any sequence, with an error of order $\frac{1}{\gamma T}$ after T steps, although coverage in a short window can still depart from the target. Foygel Barber et al. extended the analysis to non-exchangeable data and bounded the loss of coverage by a weighted sum of total variation distances, which gives a quantitative account of how much coverage drift can cost [26].

C. Risk control

Many decisions depend on a loss other than miscoverage. A tumour segmentation system is judged by the fraction of tumour pixels it misses, and a multi-label screening tool by its false negative rate. Bates et al. introduced risk-controlling prediction sets, which choose a threshold so that the expected loss falls below a target with high probability over the calibration draw [27]. Conformal risk control gives the in-expectation guarantee $E[L(\lambda)] \leq \alpha$ for any loss bounded by B and monotone in a threshold λ [10]. The threshold is the smallest λ for which:

$$\frac{n}{n+1} R_n(\lambda) + \frac{B}{n+1} \leq \alpha \tag{3}$$

where R_n is the empirical risk on the calibration data. Both guarantees hold in finite samples. Learn then test treats the selection of any set of tuning parameters as a multiple testing problem [28]. This removes the monotonicity requirement and allows several risks to be controlled together [28]. These methods carry conformal reasoning from sets to thresholded decisions in general, which is the form most high-stakes systems take.

V. APPLICATIONS AND HUMAN USE OF PREDICTION SETS

Medical imaging has received the most attention. Angelopoulos et al. derived per-pixel uncertainty intervals for image-to-image regression with a user-specified confidence guarantee and evaluated them on quantitative phase microscopy, accelerated magnetic resonance imaging and electron microscopy [29]. Olsson et al. paired an AI system for the diagnosis and grading of prostate biopsies with conformal prediction [30]. The test set held 3059 biopsies from 676 men. They report that on test data from the training laboratory and scanner, 1 of 794 predictions for cancer diagnosis was incorrect with conformal prediction, against 14 errors without it. In the same setting, 175 predictions (22%) were flagged as unreliable. The same study reports that small external samples were enough to detect systematic differences in data from other scanners and laboratories, which points to a second use of conformal output as an alarm for the distribution shift discussed in Section IV.

Language generation is a newer target. Quach et al. sample candidate responses from a language model until a calibrated stopping rule is met, so that the returned set contains at least one acceptable response with a stated probability, and their tasks include radiology report generation [31]. Mohri and Hashimoto make generated answers less specific until the remaining claims are correct with a target probability [32]. They report correctness guarantees of 80% to 90% while most of the original output is retained [32]. In both cases the guarantee concerns a set or a reduced answer defined by a scoring function, and its value depends on how well that function separates acceptable from unacceptable outputs.

A coverage guarantee helps only if the people who receive the set use it well. Straitouri et al. designed a decision support system in which the expert must choose a label from the conformal set, and simulation experiments with synthetic and real expert predictions suggested that the system can help experts predict more accurately [33]. Cresswell et al. ran a pre-registered randomised controlled trial and found, with statistical significance, that participants given conformal sets were more accurate than those given fixed-size sets with the same coverage guarantee [34]. Both studies use controlled tasks. Evidence from clinicians making real decisions with conformal sets remains scarce.

Set size is the most interpretable output of the method. A singleton is a confident answer with a known error rate, a large set signals ambiguity, and an empty set under the threshold score signals that no label fits well, which can mark an input far from the training data. The set does not explain why the model is uncertain, so it complements the interpretable modelling advocated by Rudin rather than replacing it [13]. Multimodal systems offer a further use. A reliability-weighted fusion mechanism for emotion recognition proposed in this journal is designed to estimate how trustworthy each input stream is [35], and a conformal layer over its output would turn its output scores into sets with a stated coverage.

VI. A DECISION-ORIENTED REPORTING PROTOCOL

The work reviewed above suggests that a conformal guarantee is only as useful as the description that accompanies it. This article therefore proposes a reporting protocol for conformal systems in high-stakes deployment. It has six stages, and each stage yields one or two reported items. Fig. 3 shows them and Table 2 lists the items. The protocol is untested. Its items are derived from the failure points identified in Sections III and IV.

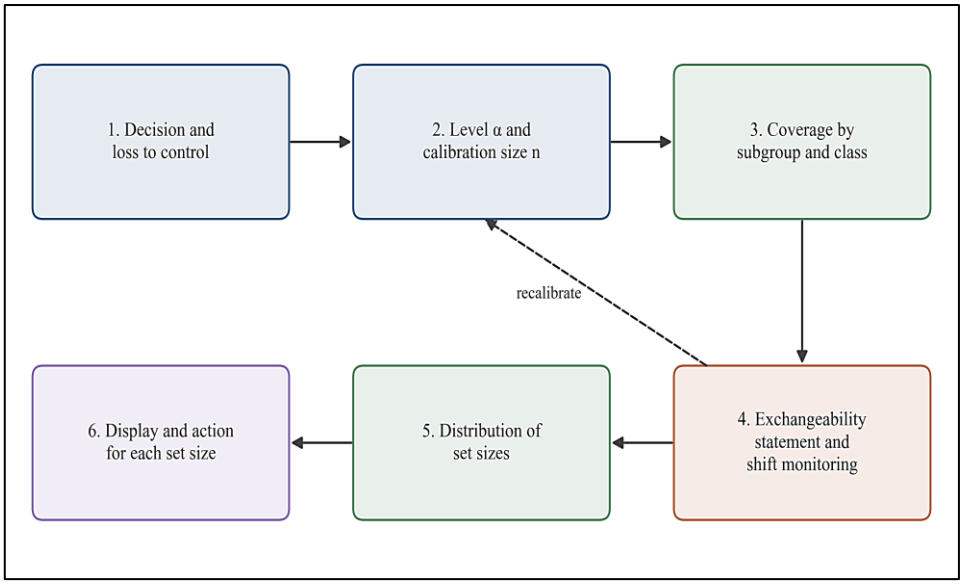


Fig. 3: Schematic of the proposed decision-oriented reporting protocol for conformal prediction in high-stakes deployment. Each stage produces a reported item listed in Table 2, and the monitoring stage feeds back to recalibration when the observed miscoverage departs from the target.

The protocol starts from the decision rather than from the model. The first stage names the action that the set supports, such as referral to a specialist whenever the set contains more than one diagnosis. It also names the loss attached to each kind of error, which determines whether miscoverage, a false negative rate or another risk should be controlled [10], [28]. The second stage fixes α with the people who bear the consequences and records the calibration size n together with the spread implied by the Beta law. A reader then sees, for example, that n = 200 at α = 0.1 places the 5th percentile of realised coverage at 0.864 (Fig. 2). The third stage reports coverage and set size for each predefined subgroup and class, with group sizes, since marginal coverage can hide failures in exactly these strata [8], [22].

The fourth stage states the exchangeability assumption in operational terms. It names the sites, devices and time period of the calibration data. It also specifies the monitoring rule that will detect shift, for example a running miscoverage rate with an adaptive update of the kind described in Section IV [25]. The fifth stage reports the distribution of set sizes rather than the average, because a mean size of 1.6 may consist of many singletons and a few very large sets. Singletons and very large sets call for different workflows. The sixth stage records how sets re shown to the user and what the user is expected to do with each size, since the benefit observed in a simulation study and a human trial depends on how the sets are built and offered [33], [34]. Table 2 pairs each item with the failure it is meant to prevent.

Table 2. Proposed reporting items for a deployed conformal system

Item	Content reported	Failure it guards against	Basis
Decision and loss	Action taken for each set outcome, loss controlled	Guarantee on the wrong quantity	[10], [28]
Level and calibration size	α, n and the 5th percentile of conditional coverage	Over-reading the nominal level	Fig. 2, [7]
Stratified coverage	Coverage and set size per subgroup and class, with counts	Hidden undercoverage in a subgroup	[23], [22]
Exchangeability statement	Sites, devices and period of the calibration data	Unnoticed shift	[9], [26]
Monitoring rule	Running miscoverage and the trigger for recalibration	Drift after deployment	[25]
Set-size distribution	Histogram of sizes, share of singletons and empty sets	Averages that hide large sets	[20]
Presentation	Display and expected user action for each size	Sets ignored or misused	[33], [34]

Three limitations apply. The protocol has not been applied to a deployed system, and its items may prove incomplete when tried in a specific clinical workflow. The values in Fig. 2 assume continuous scores and i.i.d. data, so they describe the best case rather than what a shifted deployment will achieve. The evidence on human use comes from a small number of controlled studies, and the claim that prediction sets improve expert decisions

in clinical practice is still a hypothesis. Validation would require a prospective study in which the reported items are compared with the coverage and decisions observed after deployment.

VII. CONCLUSION

Conformal prediction gives deep learning systems an output whose marginal error rate is set by the user and holds without assumptions on the model. The guarantee is marginal and depends on exchangeability. At calibration sizes of a few hundred, its realised value also spreads over several percentage points, as the Beta law in Fig. 2 shows. Extensions for group coverage, distribution shift and general risks address each of these limits in part, and each brings its own requirement for data or assumptions.

For high-stakes use, the method is most valuable when the set is tied to a named action and reported with its calibration size, its coverage by subgroup and its monitoring rule. The protocol in Table 2 is a proposal for that reporting. Testing it on real deployments, together with studies of how clinicians and other experts act on prediction sets, is the next step.

REFERENCES

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 70, 2017, pp. 1321–1330.

[2] M. Minderer *et al*., "Revisiting the calibration of modern neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 15682–15694.

[3] A. Kumar, P. Liang, and T. Ma, "Verified uncertainty calibration," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[4] Y. Ovadia *et al*., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 2nd ed. Cham, Switzerland: Springer, 2022.

[6] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, 2008.

[7] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.

[8] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Inf. Inference*, vol. 10, no. 2, pp. 455–482, 2021.

[9] R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[10] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[12] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 48, 2016, pp. 1050–1059.

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[14] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: Communicating uncertainty in medical machine learning," *npj Digit. Med.*, vol. 4, no. 1, Art. no. 4, 2021.

[15] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in *Proc. 13th Eur. Conf. Mach. Learn. (ECML)*, LNCS, vol. 2430, 2002, pp. 345–356.

[16] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Statist. Assoc.*, vol. 113, no. 523, pp. 1094–1111, 2018.

[17] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *J. Amer. Statist. Assoc.*, vol. 114, no. 525, pp. 223–234, 2019.

[18] V. Vovk, "Conditional validity of inductive conformal predictors," *Mach. Learn.*, vol. 92, nos. 2–3, pp. 349–376, 2013.

[19] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 3581–3591.

[20] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[21] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[22] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani, "Class-conditional conformal prediction with many classes," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 64555–64576.

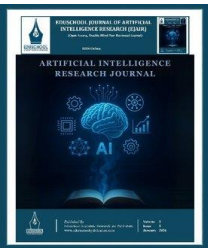
[23] Y. Romano, R. Foygel Barber, C. Sabatti, and E. Candès, "With malice toward none: Assessing uncertainty via equalized coverage," *Harvard Data Sci. Rev.*, vol. 2, no. 2, 2020.

[24] C. Lu, A. Lemay, K. Chang, K. Höbel, and J. Kalpathy-Cramer, "Fair conformal predictors for applications in medical imaging," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, 2022, pp. 12008–12016.

[25] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 1660–1672.

[26] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *Ann. Statist.*, vol. 51, no. 2, pp. 816–845, 2023.
[27] S. Bates, A. N. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, "Distribution-free, risk-controlling prediction sets," *J. ACM*, vol. 68, no. 6, pp. 1–34, 2021.
[28] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *Ann. Appl. Statist.*, vol. 19, no. 2, pp. 1641–1662, 2025.
[29] A. N. Angelopoulos *et al*., "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 162, 2022, pp. 717–730.
[30] H. Olsson *et al*., "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nat. Commun.*, vol. 13, Art. no. 7761, 2022.
[31] V. Quach *et al*., "Conformal language modeling," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
[32] C. Mohri and T. Hashimoto, "Language models with conformal factuality guarantees," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 235, 2024, pp. 36029–36047.
[33] E. Straitouri, L. Wang, N. Okati, and M. Gomez Rodriguez, "Improving expert predictions with conformal prediction," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 202, 2023, pp. 32633–32653.
[34] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis, "Conformal prediction sets improve human decision making," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 235, 2024, pp. 9439–9457.
[35] N. Chandran, "Audio-visual-textual fusion for emotion recognition: A neurologically inspired attention mechanism," *Eduschool J. Artif. Intell. Res.*, vol. 1, no. 2, pp. 8–14, 2026, doi: 10.63090/EJAIR/3139.7336.0007.

EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Cross-Modal Retrieval with Modality-Agnostic Embeddings Towards Universal Representation Learning

Rejina P V

Assistant professor, Co-Operative Arts and Science College, Madayi , Pazhayangadi, Kannur, India

Article information

Received: 3rd August 2026
Received in revised form: 5th September 2026
Accepted: 7th October 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0010

Abstract

A single vector space that accepts images, text, audio, video and sensor readings, and that supports retrieval between any pair of them, has become an explicit goal of multimodal research. Systems now exist that approach it by binding every modality to one anchor, which produces retrieval between pairs that were never trained together. This article examines what such spaces achieve and where they fail. We first analyse the alignment topologies available, showing that binding through an anchor reduces the requirement for paired data from a quadratic to a linear function of the number of modalities, and that the reduction has a measurable cost in the quality of pairings the anchor does not mediate. We then examine the geometry of the resulting space. Published analysis has shown that representations of different modalities occupy separate regions rather than mixing, a separation that persists throughout training and that has consequences for retrieval which are widely overlooked. Hubness and anisotropy compound the effect at the scale required for practical search. We compare seven systems by their binding topology and coverage, review the approximate search structures that make retrieval feasible over billions of items, and assess the claim that models trained on different modalities are converging on a shared statistical description of the world. Six research problems are identified.

Keywords:- Cross-Modal Retrieval, Modality-Agnostic Embeddings, Modality Gap, Binding Topology, Hubness, Universal Representation

.

I. INTRODUCTION

Retrieval across modalities was once a matter of building a bridge between two specific representations, with a separate bridge for each pair of interest. Contrastive pre-training on image and caption data replaced the bridge with a shared space in which both inputs are encoded as vectors of equal dimension and compared by inner product [1], [2]. The construction generalises. If audio is aligned to images, and text is aligned to images, then audio and text can be compared through the space even though no audio and text pair was ever presented during training. Systems built on this observation report retrieval between such pairs and describe the capability as emergent [3], [4].

The appeal is practical as well as conceptual. A single index serves every query type, a new modality can be added by training one encoder against an existing anchor, and downstream systems need to handle only one kind of vector. The appeal has carried the idea into deployment, where shared embedding spaces now support search over mixed collections of photographs, documents and recordings. What has received less attention is the internal structure of these spaces and whether the geometry supports the uses to which they are being put.

It does not, at least not without correction. Analysis of contrastive spaces has established that embeddings of different modalities occupy distinct regions separated by a substantial distance, that the separation is present at initialisation and is preserved rather than removed by training, and that it affects downstream behaviour [8]. A space in which every image lies in one region and every caption in another is not modality-agnostic in any meaningful sense, and the fact that retrieval works despite this reflects the local ordering of similarities rather than any genuine mixing. Two further properties of high dimensional spaces, namely the emergence of hub points that appear among the nearest neighbours of a large number of queries [9] and the concentration of embeddings within a narrow cone, compound the difficulty at the scale of a practical index.

This article treats the shared space as an object of study rather than as a means to an end. Section II compares alignment topologies and quantifies the paired data each requires. Section III examines the geometry of the resulting space and the three properties that govern retrieval quality. Section IV compares published systems and reviews the search structures needed at scale. Section V assesses the claim that representations from different modalities are converging on a common description of the world, and states six open problems.

II. ALIGNMENT TOPOLOGIES

Three arrangements are possible when more than two modalities are involved, and Fig. 1 shows them. Pairwise alignment trains a separate objective for every pair, which requires paired data for each and produces the strongest correspondence for the pairs that receive it. Anchor binding aligns every modality to one designated modality, usually images or text, and relies on transitivity for the rest. Joint encoding maps all modalities through a single network with modality-specific input adapters, so that alignment is a consequence of shared parameters rather than of an explicit pairwise objective [10], [11], [12].

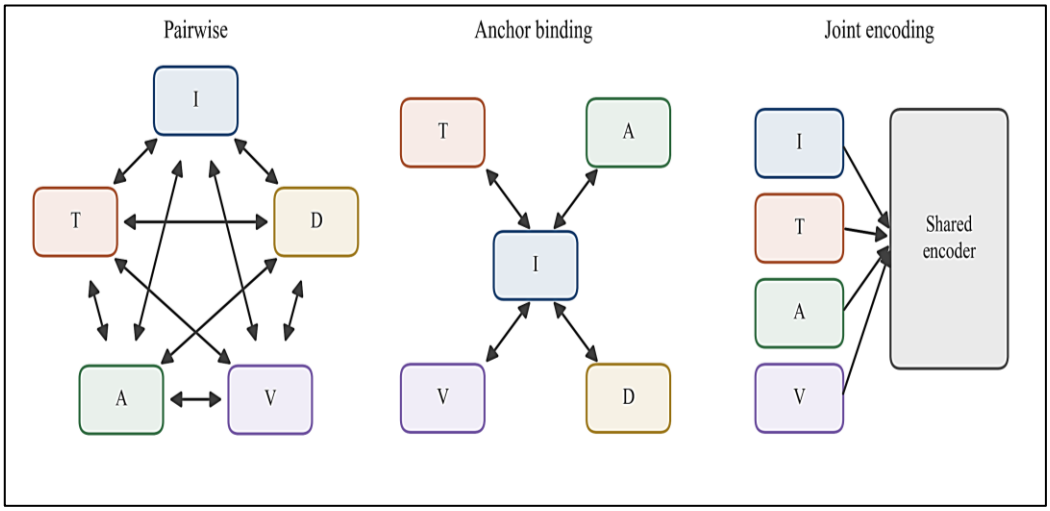


Fig. 1: Three alignment topologies for a shared embedding space over five modalities. Pairwise alignment requires paired data for every edge, anchor binding requires it only for edges incident on the anchor, and joint encoding replaces explicit pairing with shared parameters and modality-specific adapters.

The difference in data requirement is decisive. Pairwise alignment over N modalities requires paired corpora for N(N-1)/2 combinations, while anchor binding requires N-1. Fig. 2 plots both, and the divergence is severe by the time a system covers six or seven signal types.

The practical position is worse than the counting suggests, because paired corpora do not exist in comparable quantity for most combinations. Image and text pairs are available in billions [17], audio and text pairs in the tens of thousands [20], [30], and pairs joining depth or inertial measurement to language are scarce. Anchor binding is therefore not only cheaper but is the only arrangement for which the data exist, which is the reason it has become standard [3], [4].

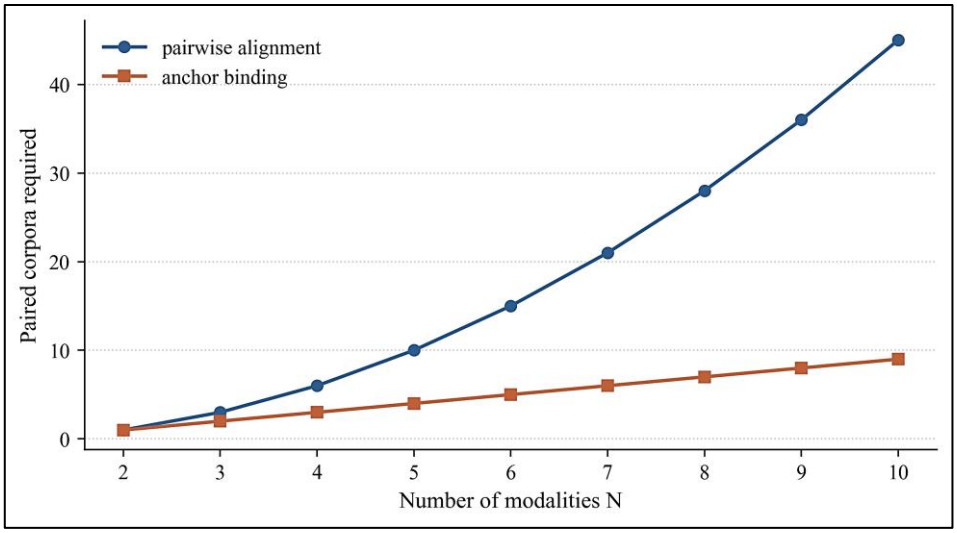


Fig. 2: Number of paired corpora required to align N modalities under pairwise alignment and under anchor binding. Values are computed from the combinatorial expressions N(N-1)/2 and N-1 respectively.

The cost of the reduction is that transitivity is not exact. Alignment of audio to images preserves the information the two share, and alignment of text to images preserves a different overlap. What survives the composition of the two is the intersection, so retrieval between audio and text through an image anchor is restricted to content that images can express. A sound that has no visual correlate, such as the tone of an utterance or a mechanical fault audible in a running engine, has no representation the anchor can carry. Reported results for emergent pairs are consistent with this account, since the accuracy attained on pairs mediated by an anchor is below that attained on directly trained pairs even when the encoders are the same [3]. Table 1 sets out the properties of the three topologies.

Table 1. Properties of the three alignment topologies

Topology	Paired corpora needed	Emergent pairs	Principal advantage	Principal limitation
Pairwise	N(N-1)/2	Not required	Strongest correspondence per pair	Data do not exist for most pairs
Anchor binding	N-1	Yes, through the anchor	New modality added by one training run	Content absent from the anchor cannot transfer
Joint encoding	Varies with supervision	Yes, through shared parameters	One network for all inputs	Capacity shared across dissimilar signals

The encoders are largely common across the three topologies, which is worth stating because it locates the differences in the objective rather than in the network. Almost every system uses transformer blocks [29] with patch or token embeddings supplying the input, following the formulation established for images [28], and almost every alignment objective descends from noise contrastive estimation applied to a batch of candidates [16]. Designs that retain a fusion encoder rather than producing independent vectors achieve finer alignment and are used for reranking a shortlist [27], [32], and the two modality retrieval systems built for video follow the same pattern with temporal pooling added [25], [26]. Unified masked modelling over image and text tokens reaches comparable quality with a single objective [22]. What separates the modality-agnostic systems from these is not the encoder but the decision to align many modalities into one space, and the properties examined in the next section are consequences of that decision.

III. THE GEOMETRY OF SHARED SPACES

A. The modality gap

The most consequential property of contrastive embedding spaces is that they do not mix. Embeddings of images occupy one region of the hypersphere and embeddings of text another, with a clear separation between them, and the separation is not an artefact of insufficient training. It arises from the narrow cone in which randomly

initialised deep networks place their outputs, combined with a contrastive objective that preserves the initial separation because moving the two groups together is not required to rank the correct pair first [8]. Fig. 3 illustrates the arrangement on the unit circle. The correct pair is closer than any incorrect pair, so retrieval succeeds, yet no image embedding is close to any text embedding in absolute terms.

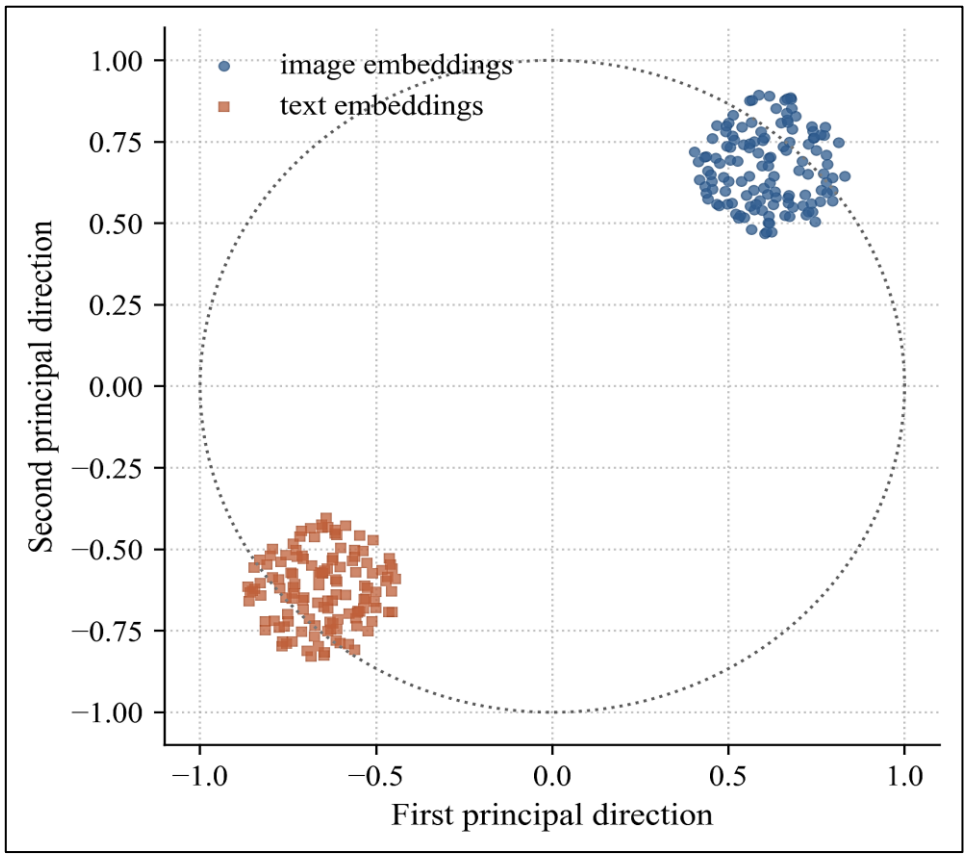


Fig. 3: Schematic illustration of the modality gap on the unit sphere, drawn from a fixed seed. Embeddings of the two modalities occupy separate regions while the ranking within each region supports retrieval. The arrangement follows the analysis reported in [8].

Three consequences follow for retrieval. Scores are not comparable across modality pairs, so a threshold calibrated for image to text search is wrong for audio to text search, and any system that fuses results from several pair types must recalibrate. Arithmetic in the space is unreliable, since adding a text vector to an image vector produces a point in the empty region between the two groups where no encoder places anything. Fine tuning that moves the groups together can improve or degrade downstream accuracy depending on the task, which means the gap is not simply a defect to be removed [8]. Practical remedies include centring each modality by subtracting its mean before comparison, whitening within modality, and training with a formulation that scores each pair independently rather than through a softmax over the batch [21].

B. Hubness and anisotropy

Two further properties appear at scale. Hubness is the tendency in high dimensional spaces for a small number of points to appear in the neighbour lists of a disproportionate share of queries, a phenomenon established for general vector data and directly relevant to retrieval [9]. In a cross-modal index, hub items are returned for queries they do not match, which lowers precision in a way that averages conceal because the affected queries are a minority. Corrections that normalise similarity by the average similarity of each candidate to a sample of queries are effective and inexpensive, and their absence from most published retrieval pipelines is an oversight rather than a considered choice.

Anisotropy is the concentration of embeddings within a narrow cone, which compresses the range of achievable similarity scores and reduces the separation between a correct match and a plausible distractor. The effect interacts with quantisation, since a compression scheme that allocates codewords uniformly over the sphere wastes capacity on regions no embedding occupies [15]. Whitening or applying a learned rotation before quantisation recovers part of the loss and is standard practice in text retrieval, though it is applied inconsistently in the cross-modal setting.

IV. SYSTEMS AND RETRIEVAL AT SCALE

Table 2 compares seven systems that construct shared spaces across more than two modalities. The pattern is consistent, with images or text serving as the anchor and additional modalities attached by contrastive objectives against the anchor. Coverage has grown from three modalities to six, and the newer systems use language rather than images as the anchor on the argument that language expresses more of what the other modalities carry [4]. Systems that avoid the anchor entirely by encoding all inputs through shared parameters have demonstrated broad task coverage [10], [11], [12], [23], though their retrieval accuracy on standard cross-modal benchmarks has not matched that of dedicated contrastive models.

Table 2. Systems constructing shared embedding spaces over multiple modalities

System	Year	Modalities covered	Anchor	Alignment objective	Ref.
CLIP	2021	Image, text	Not applicable	Symmetric InfoNCE	[1]
VATT	2021	Video, audio, text	Video	Hierarchical contrastive	[6]
AudioCLIP	2022	Image, text, audio	Image and text	Contrastive extension	[5]
CLAP	2023	Audio, text	Not applicable	Contrastive	[7], [24]
ImageBind	2023	Image, text, audio, depth, thermal, inertial	Image	Pairwise contrastive to anchor	[3]
LanguageBind	2024	Video, infrared, depth, audio, text	Language	Pairwise contrastive to anchor	[4]
Meta-Transformer	2023	Twelve input types	Not applicable	Shared encoder with adapters	[12]

Evaluation of these systems relies on benchmarks built for two modalities, principally curated image and caption sets [18], [19] and audio caption collections [20]. No benchmark measures the property that distinguishes a modality-agnostic space, namely consistency of the geometry across all pairs. A system could score well on every pair separately while placing the modalities in mutually inconsistent arrangements, and no reported protocol would detect it. A suitable measure would compare the similarity assigned to a triple of items describing the same content across three modalities and report the variance of those similarities, which would test transitivity directly.

Retrieval at deployment scale depends on approximate search. Exhaustive comparison against a billion vectors is infeasible for interactive use, so systems use graph based indexes that traverse a hierarchy of neighbour links [14] or partitioned indexes with compressed residuals, with product quantisation the standard compression [15], and implementations on parallel hardware make billion scale search practical [13]. The interaction between these structures and the geometry described in Section III is the part most often ignored. A graph index built over embeddings that occupy two separated regions will place few edges between the regions, so a query from one modality enters the graph far from its target and the search must cross the gap before it can descend. Building separate indexes per modality and centring each one removes the problem and is the recommended arrangement.

V. TOWARDS UNIVERSAL REPRESENTATIONS

The strongest form of the universal representation claim is that models trained on different data and different modalities converge on the same underlying description of the world, so that alignment amounts to discovering a rotation between spaces that already agree. Evidence has been offered for this position in the form of measurements showing that representational similarity between independently trained models increases with their scale and performance, across vision and language alike [34]. The analogy with unsupervised alignment of word embeddings between languages is instructive, since translation without parallel data was achieved by finding a linear map between separately trained spaces on the basis of their shared structure [33].

The evidence supports a weaker reading than the strong claim requires. Convergence is measured between models trained on data that overlap in content, and the shared structure may reflect a shared world rather than a shared representational format. The modality gap is direct evidence against complete convergence, since spaces that agreed would not remain separated after alignment. A reasonable position is that a common core exists for content that several modalities can express, that this core is what anchor binding transfers, and that each modality retains information outside it which no shared space will carry.

Six problems follow. The first is a measure of transitivity, without which the central property of a modality-agnostic space is unmeasured. The second is representation of modality-specific content, which a single shared

vector cannot hold and which a composite representation with a shared and a private component could. The third is calibration of similarity across pairs, so that scores from different modality pairs can be compared and combined. The fourth is the treatment of hubness and anisotropy as part of the retrieval pipeline rather than as post hoc corrections. The fifth is the addition of modalities without retraining, since the anchor arrangement requires a new contrastive run for each and cheaper adaptation methods have been demonstrated for the two modality case [31]. The sixth is evaluation on modalities with weak visual correlates, such as inertial measurement and thermal imaging, where anchor binding through images is least likely to transfer and where reported results are thinnest [3].

VI. CONCLUSION

Shared embedding spaces have made cross-modal retrieval a matter of encoding rather than of engineering, and binding through an anchor has reduced the paired data requirement from a quadratic to a linear function of the number of modalities, which is what makes coverage of six or more signal types feasible. The arrangement carries a cost that is intrinsic rather than incidental, since only content the anchor can express passes between the modalities it mediates.

The geometry of these spaces deserves more attention than it receives. Modality separation persists through training and affects score calibration, vector arithmetic and index construction. Hubness and anisotropy degrade precision at the scale at which these systems are deployed, and inexpensive corrections exist but are inconsistently applied. The claim that independently trained models are converging on a universal representation is supported for content that several modalities share and is contradicted by the separation the same spaces exhibit. Progress requires a measure of transitivity, a representation that admits modality-specific content alongside the shared component, and evaluation on signals for which no visual anchor is available.

REFERENCES

[1] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[2] C. Jia et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 4904–4916.

[3] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “ImageBind: One embedding space to bind them all,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 15180–15190.

[4] B. Zhu et al., “LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[5] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “AudioCLIP: Extending CLIP to image, text and audio,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 976–980.

[6] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, “VATT: Transformers for multimodal self-supervised learning from raw video, audio and text,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 24206–24221.

[7] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “CLAP: Learning audio concepts from natural language supervision,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[8] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 17612–17625.

[9] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, “Hubs in space: Popular nearest neighbors in high-dimensional data,” *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[10] A. Jaegle et al., “Perceiver IO: A general architecture for structured inputs and outputs,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[11] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, “Unified-IO: A unified model for vision, language, and multi-modal tasks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[12] Y. Zhang et al., “Meta-Transformer: A unified framework for multimodal learning,” *arXiv preprint* arXiv:2307.10802, 2023.

[13] J. Johnson, M. Douze, and H. Jegou, “Billion-scale similarity search with GPUs,” *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[14] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, 2020.

[15] H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, 2011.

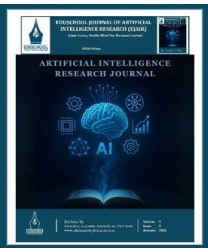
[16] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint* arXiv:1807.03748, 2018.

[17] C. Schuhmann et al., “LAION-5B: An open large-scale dataset for training next generation image-text models,” in *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*, 2022.

[18] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[19] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2641–2649.
[20] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2019, pp. 119–132.
[21] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 11975–11986.
[22] W. Wang et al., "Image as a foreign language: BEiT pretraining for vision and vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 19175–19186.
[23] S. Reed et al., "A generalist agent," *Trans. Mach. Learn. Res.*, 2022.
[24] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.
[25] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 6787–6800.
[26] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1728–1738.
[27] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5583–5594.
[28] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[29] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
[30] D. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 736–740.
[31] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
[32] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19730–19742.
[33] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jegou, "Word translation without parallel data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
[34] M. Huh, B. Cheung, T. Wang, and P. Isola, "The platonic representation hypothesis," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024.

EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Grounded Language Generation from Raw Sensory Input in Embodied AI Systems

Manasy Jayasurya

Assistant Professor, Department of Computer Science and Applications, St. Mary's College (Autonomous), Thrissur, India

Article information

Received: 5th August 2026
Received in revised form: 8th September 2026
Accepted: 10th October 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0011

Abstract

Language produced by a system that has never acted in the world describes that world from the outside. Embodied systems that perceive through their own sensors and change what they perceive by acting offer a different setting for language generation, one in which the meaning of a term can be tied to sensory evidence and to the actions the term licenses. This article examines what such grounding requires and how far current systems supply it. We distinguish three requirements, namely perceptual anchoring of reference, action conditioned meaning and situated deixis, and show that pipelines which caption an image and pass the caption to a planner satisfy only the first. End-to-end architectures that map sensory observations and instructions directly to actions and to language remove the intermediate text bottleneck, and we compare eight such systems by input, output and training signal. We then examine the data available, which is smaller than the corpora used for text pre-training by five orders of magnitude, and the evaluation problem, where reference based caption metrics reward fluency and cannot detect a description of an object that is not present. A taxonomy of grounding failures is proposed together with the diagnostic each one requires, and seven open problems are stated.

Keywords:- Symbol Grounding, Embodied Artificial Intelligence, Vision-Language-Action Models, Situated Language Generation, Robot Learning, Hallucination

.

I. INTRODUCTION

A system trained only on text learns the relations between words and other words. The question of whether such relations amount to meaning was posed precisely as the symbol grounding problem, which asks how the symbols manipulated by a formal system acquire any connection to what they are about other than through further symbols [1]. Work in cognitive science supplies one answer, namely that concepts are represented in the same systems that support perception and action, so that understanding a word about grasping involves the machinery used for grasping [2]. A position paper directed at the language processing community made the argument in operational form, holding that the scope of what a system can learn is bounded by the experience available to it, and that text alone occupies a narrow position on that scale [3]. Embodied systems change the terms of the problem. A robot that perceives through cameras and force sensors, and that alters its own observations by moving, has access to the correspondence between a word and an experience. The correspondence can be learned rather than stipulated, which was demonstrated in early work that acquired word forms and their visual referents from recordings of caregivers speaking to infants [39], and pursued further in systems that connected instructions to perceived objects and navigable routes [38], [40]. The scale available now is different, but the requirement is

unchanged. Two developments have made the topic active again. Large vision and language models trained on web data provide a starting point with broad recognition ability and fluent generation [17], [18], [19], and models that map images and instructions directly to robot actions have shown that this ability transfers to control [10], [11], [12], [14]. Yet the language these systems produce remains weakly tied to what their sensors report. Vision and language models describe objects that are not present at rates measurable on purpose built diagnostics [21], [22], and a system that misdescribes a scene it is acting in fails in a way that a captioning system does not, because the description informs the next action.This article examines grounded generation as a problem in its own right rather than as a component of control. Section II states three requirements for grounding and separates them from recognition. Section III compares modular and end-to-end architectures and reviews eight systems. Section IV examines the training data, whose scale is the binding constraint. Section V addresses evaluation, proposes a taxonomy of grounding failures and identifies the diagnostic each requires. Section VI states seven open problems and Section VII concludes.

II. WHAT GROUNDING REQUIRES

Grounding is used loosely in current writing, and separating its components clarifies what a given system achieves. The first requirement is perceptual anchoring, meaning that a referring expression corresponds to evidence in the current sensory input rather than to a prior over what is usually present. A system satisfies this requirement when its description of an object changes appropriately as the object changes, and fails it when the description reflects the statistics of the training corpus. Object hallucination measurements target exactly this failure, and reported rates on constructed probes remain substantial for systems that score well on caption benchmarks [21], [22].The second requirement is action conditioned meaning. Terms such as reachable, heavy, blocked and open describe relations between an object and the capabilities of the agent, and no observer without a body can verify them. A system that has attempted to lift objects has evidence about weight that no amount of image and caption data supplies. Affordance grounded language selection makes the requirement operational by scoring candidate instructions with a value function that estimates whether the described action can be executed from the current state, so that a plan is restricted to what the agent can actually do [9]. The third requirement is situated deixis, covering terms whose referent depends on the position and history of the speaker, including here, this one, behind you and the one I just put down. These cannot be resolved from a single frame, and they are the terms most common in natural instruction, which is why dialogue based benchmarks include them explicitly [27], [28]. The three requirements are ordered by the machinery they demand. Perceptual anchoring needs an encoder and a mechanism that ties generation to it. Action conditioned meaning needs a policy or a value function and the experience to fit it. Situated deixis needs persistent state covering the agent position and the recent history of the interaction. Almost all reported progress addresses the first requirement, which is also the one that transfers most readily from web data, and the third is addressed only in benchmarks designed for it.

III. ARCHITECTURES FOR GROUNDED GENERATION

Two designs are in use, and Fig. 1 contrasts them. The modular pipeline runs a perception module that produces text, passes the text to a language model that plans, and passes the resulting instruction to a controller. The arrangement is attractive because each component can be developed separately and the intermediate text is readable, and it has produced working systems for both planning [31] and composed reasoning across pre-trained models [32]. Its defect is the bottleneck at the text interface, since the caption discards everything the planner was not expected to need, including spatial detail, uncertainty and the appearance of objects that no caption template covers. When the planner asks a question the caption cannot answer, there is no path back to the sensor.

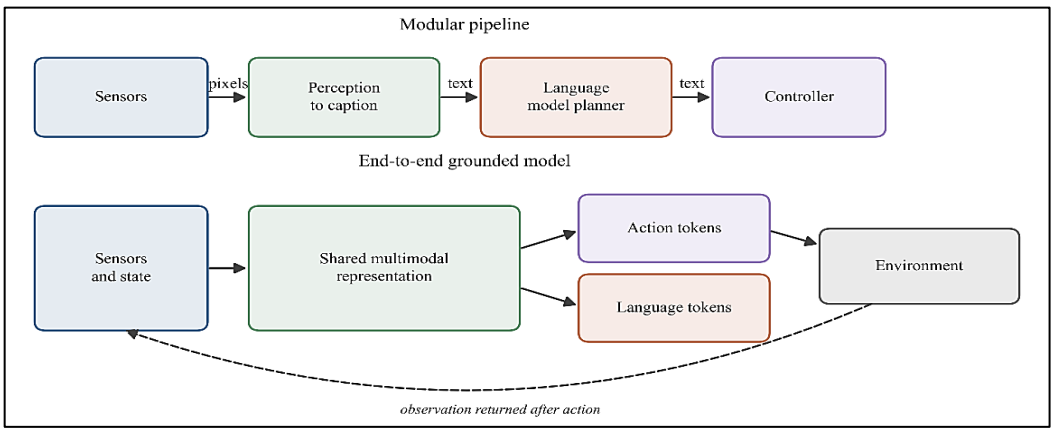


Fig. 1: Modular and end-to-end designs for grounded language generation. The modular pipeline passes information between components as text, which discards sensory detail. The end-to-end design conditions both action and language on a shared representation of observation, instruction and state.

The end-to-end design encodes observations, instruction and state into a shared representation from which both actions and language are produced. Robotics transformers established that actions can be emitted as discrete tokens by a sequence model trained on demonstration data [10], and the vision-language-action formulation extended this by initialising from a model pre-trained on web image and text data, so that the semantic knowledge acquired from the web is available for control [11]. An embodied multimodal language model injects continuous sensor embeddings directly into the token stream of a language model, which allows the same network to plan, answer questions and describe its situation without an intermediate caption [12]. Open implementations of the vision-language-action recipe have made the approach reproducible [14], and generalist sequence models trained across many tasks and embodiments show that a single set of parameters can serve varied outputs [13].

Between the two designs lies a family that keeps a pre-trained visual and linguistic representation and attaches a spatial mechanism to it. Manipulation policies that combine a semantic pathway from a contrastive vision-language model with a spatial pathway that predicts where to act follow this pattern [15], and methods that compose value maps in three dimensions from code produced by a language model achieve similar effects without task specific training [16]. Table 1 compares eight systems along the dimensions that matter for grounded generation, namely what the model consumes, what it emits, where its training signal comes from and whether it produces language at all during operation.

Table 1. Systems for embodied grounding compared by input, output and training signal

System	Year	Input	Output	Training signal	Emits language	Ref.
Gato	2022	Images, text, proprioception	Actions and text	Supervised across tasks	Yes	[13]
SayCan	2022	Image, instruction	Skill selection	Value function plus language model	Yes	[9]
CLIPort	2022	Image, instruction	Pick and place poses	Demonstrations	No	[15]
RT-1	2023	Image history, instruction	Discretised actions	Demonstrations	No	[10]
PaLM-E	2023	Sensor embeddings, text	Text and plans	Mixed robot and web data	Yes	[12]
RT-2	2023	Image, instruction	Action tokens	Web pre-training plus demonstrations	Partly	[11]
VoxPoser	2023	RGB-D, instruction	Value maps for planning	No task specific training	Yes	[16]
OpenVLA	2024	Image, instruction	Action tokens	Open cross-embodiment data	No	[14]

The final column of Table 1 identifies the gap this article addresses. Most systems that act do not speak, and most systems that speak do not act. The two that do both produce language conditioned on sensory input [12], [13], and neither has been evaluated for the accuracy of that language against what the sensors actually contained. A system that acts correctly while describing the scene incorrectly is a common and largely unmeasured outcome, and it matters wherever the description is presented to a person as an account of what the system is doing.

IV. THE TRAINING SIGNAL AND ITS SCALE

The constraint on embodied grounding is data. Fig. 2 shows the scale of the principal resources, and the comparison with text pre-training is instructive, since a corpus of a trillion tokens exceeds the largest robot demonstration collection by roughly five orders of magnitude when both are measured in units of experience. Simulated environments supply volume at the cost of realism, with photorealistic indoor scans supporting navigation and instruction following [6], [8] and interactive simulators supporting manipulation of household objects [7]. Instruction following benchmarks built on these environments provide paired language and trajectory data at the scale of tens of thousands of episodes [4], [5]. Real robot data are scarcer, and the response has been aggregation, with a collaborative collection joining datasets from many laboratories and embodiments into one resource [33].

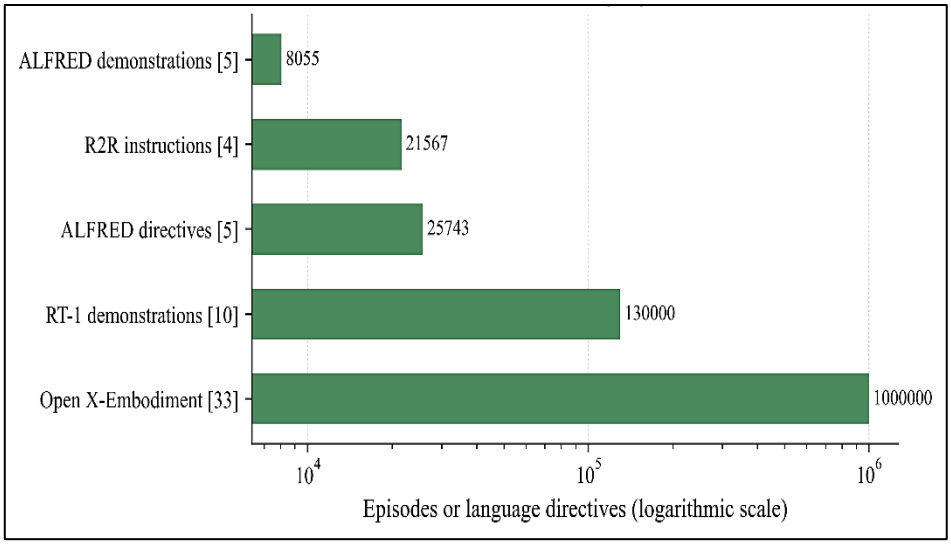


Fig. 2: Scale of paired language and experience resources for embodied learning, in episodes or instructions on a logarithmic scale, as reported by the authors of each resource: room-to-room navigation instructions [4], household task directives [5], real robot demonstrations [10] and the aggregated cross-embodiment collection [33].

Two responses to the shortage are visible in current work. The first is transfer from web data, on the argument that recognition and commonsense knowledge acquired from images and text carry to embodied settings and that only the mapping to action must be learned from robot data. Reported results support this for object identification and for instruction interpretation [11], [12], and they do not establish transfer for the action conditioned terms of Section II, which have no counterpart in web text. The second response is egocentric recording at scale, where thousands of hours of first person video capture human activity with narration, providing experience that is embodied even though the recorder was not an agent under the system's control [34].

Neither response addresses the deeper asymmetry. Language data are abundant and passive, while action data are scarce and active, and the terms that most require grounding are those tied to action. A system can learn from a billion captions what a cup looks like and cannot learn from any of them how heavy a full one is. World models offer a partial route, since a model that predicts future observations conditioned on actions can be trained from unlabelled interaction and can supply the counterfactual experience that language about affordances requires [36], [37]. Multimodal prompting, where a task is specified by a mixture of text and images rather than by text alone, reduces the amount of language that must carry the specification [35], which is a practical mitigation rather than a solution.

V. EVALUATION AND A TAXONOMY OF GROUNDING FAILURE

Caption evaluation was designed for a setting in which several human written references are available and the task is to produce something similar. Overlap based measures reward fluency and phrasing typical of the reference set [24], consensus weighted measures improve the correlation with human judgment [23], and scene graph based measures compare propositions rather than strings [25]. All of them share a defect for the present purpose, which is that a description mentioning an object absent from the scene may still score well when the rest of the sentence matches the references. Reference free measures that score the description against the image with a contrastive model avoid the dependence on references [26] and inherit the weaknesses of the model they use, including the compositional insensitivity documented for such models.

Fig. 3 sets out four failure modes and the diagnostic each requires. Fabricated reference is the description of an entity with no sensory support, which purpose built probes detect by asking about the presence of objects selected according to their frequency in the training data [21], [22]. Frozen reference is a description that remains correct for an earlier state and is not updated as the scene changes, which is detected by presenting a sequence in which one attribute changes and measuring the lag. Capability misattribution is the assertion that an action is possible when the agent cannot perform it, detected by comparing generated claims against the value function or against execution outcomes [9]. Deictic failure is the incorrect resolution of a term whose referent depends on position or history, detected in dialogue settings where the referent is established in an earlier turn [27], [28].

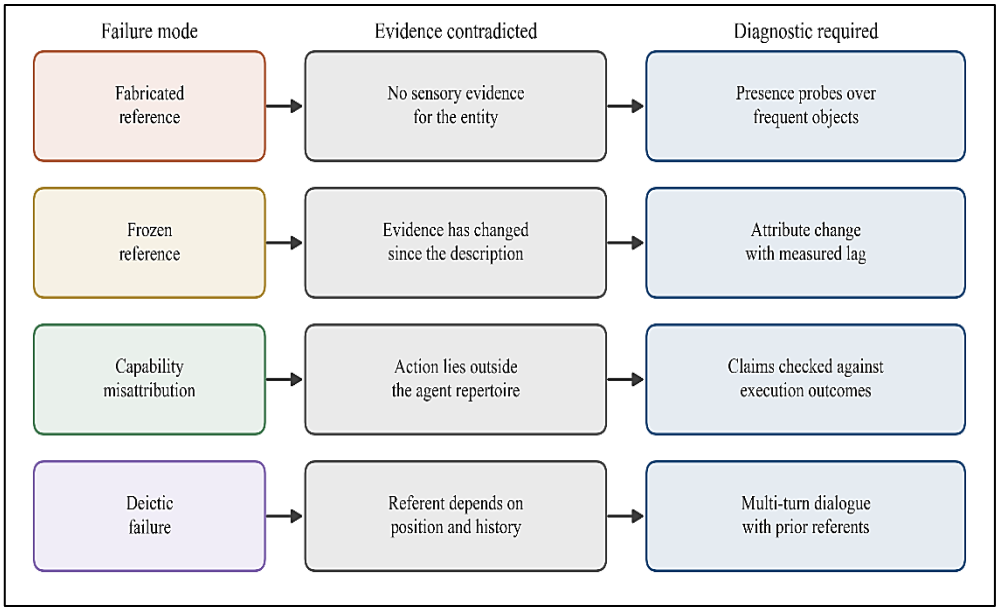


Fig. 3: Four modes of grounding failure in embodied language generation, with the sensory or behavioural evidence each contradicts and the diagnostic procedure required to detect it.

The taxonomy has a practical use beyond classification. Each mode calls for a different remedy. Fabricated reference responds to training objectives that penalise unsupported statements and to decoding constrained by detected entities. Frozen reference responds to architectural changes that force recomputation from current observations rather than reuse of a cached representation. Capability misattribution responds to the affordance scoring already demonstrated for action selection [9], applied to generation rather than to planning. Deictic failure responds to explicit state, meaning a record of agent position and of referents introduced in earlier turns. Treating all four as one phenomenon under the heading of hallucination, as much of the current literature does [20], obscures these distinctions and leads to remedies aimed at the wrong mode.

VI. OPEN PROBLEMS

Seven problems follow from the analysis. The first is an evaluation suite for embodied description that reports the four failure modes separately rather than a single accuracy figure. The second is a training objective in which the correctness of generated language is checked against the environment, which is available in simulation where a scene graph exists and which would supply the supervision that caption corpora cannot [6], [7], [29]. The third is action conditioned vocabulary, requiring corpora in which terms for weight, reachability and obstruction are paired with the outcomes of attempts rather than with images.

The fourth is persistent situated state, without which deictic terms cannot be resolved and which no current architecture maintains explicitly. The fifth is the latency budget, since a model large enough to generate fluent language is too slow for a control loop, and the two rate hierarchy suggested by current systems, with a fast policy and a slower language process, needs a principled treatment. The sixth is transfer measurement, distinguishing what web pre-training genuinely contributes from what it merely appears to contribute through benchmark overlap [11], [12], [30]. The seventh is safety in description, since a system whose account of its own actions is unreliable cannot be supervised by a person reading that account, and the requirement is more stringent than for a captioning system because the consequences are physical.

VII. CONCLUSION

Grounding is not one property. Perceptual anchoring, action conditioned meaning and situated deixis make different demands, and current systems supply the first well, the second partly and the third almost not at all. The architectures that act do not generally describe, the architectures that describe have not been evaluated on whether their descriptions match what their sensors reported, and the data available for the terms that most require embodiment remain smaller than text corpora by several orders of magnitude.

The evaluation position is the most immediately addressable. Caption metrics measure similarity to reference text and cannot detect a fluent description of a scene the system is not in. Separating grounding failure into fabricated reference, frozen reference, capability misattribution and deictic failure identifies four distinct diagnostics, each of which can be built from existing simulators and benchmarks. Beyond evaluation, the substantive requirements are objectives that check generated language against the environment, corpora that pair

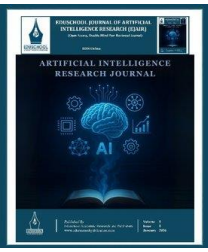
action outcomes with the vocabulary describing them, and architectures that maintain the situated state on which the most ordinary instructions depend.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, nos. 1–3, pp. 335–346, 1990.
[2] L. W. Barsalou, "Grounded cognition," *Annual Review of Psychology*, vol. 59, pp. 617–645, 2008.
[3] Y. Bisk *et al.*, "Experience grounds language," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 8718–8735.
[4] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3674–3683.
[5] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "ALFRED: A benchmark for interpreting grounded instructions for everyday tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10740–10749.
[6] M. Savva *et al.*, "Habitat: A platform for embodied AI research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9339–9347.
[7] E. Kolve *et al.*, "AI2-THOR: An interactive 3D environment for visual AI," *arXiv preprint arXiv:1712.05474*, 2017.
[8] A. Chang *et al.*, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vision (3DV)*, 2017, pp. 667–676.
[9] M. Ahn *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," in *Proc. Conf. Robot Learning (CoRL)*, 2022, pp. 287–318.
[10] A. Brohan *et al.*, "RT-1: Robotics transformer for real-world control at scale," in *Proc. Robotics: Science and Systems (RSS)*, 2023.
[11] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. Conf. Robot Learning (CoRL)*, 2023, pp. 2165–2183.
[12] D. Driess *et al.*, "PaLM-E: An embodied multimodal language model," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 8469–8488.
[13] S. Reed *et al.*, "A generalist agent," *Trans. Mach. Learn. Res.*, 2022.
[14] M. J. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," in *Proc. Conf. Robot Learning (CoRL)*, 2024.
[15] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learning (CoRL)*, 2022, pp. 894–906.
[16] W. Huang *et al.*, "VoxPoser: Composable 3D value maps for robotic manipulation with language models," in *Proc. Conf. Robot Learning (CoRL)*, 2023, pp. 540–562.
[17] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 34892–34916.
[19] J.-B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 23716–23736.
[20] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
[21] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 4035–4045.
[22] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023, pp. 292–305.
[23] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.
[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.
[25] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 382–398.
[26] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 7514–7528.
[27] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "TEACh: Task-driven embodied agents that chat," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2017–2025.
[28] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Proc. Conf. Robot Learning (CoRL)*, 2019, pp. 394–406.
[29] R. Krishna *et al.*, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[30] T. Brown *et al.*, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[31] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 9118–9147.
[32] A. Zeng *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[33] [A. Padalkar *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 6892–6903.
[34] [34] K. Grauman *et al.*, "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18995–19012.
[35] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "VIMA: General robot manipulation with multimodal prompts," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 14975–15022.
[36] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.
[37] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 2450–2462.
[38] [38] R. J. Mooney, "Learning to connect language and perception," in *Proc. AAAI Conf. Artif. Intell.*, 2008, pp. 1598–1601.
[39] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.
[40] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 1507–1514.

EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Scalable Oversight for Advanced AI: Recursive Reward Modeling and Debate

Raji N

Assistant Professor, Department of Computer Science, YuvakshatraInstitute of Management Studies (YIMS), Mundur, India

Article information

Received: 10th August 2026
Received in revised form: 12th September 2026
Accepted: 14th October 2026
Available online: 20th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0012

Abstract

Reinforcement learning from human feedback trains language models against a reward model fitted to human comparisons, and it works only as long as the people giving feedback can tell a good output from a bad one. Scalable oversight is the problem of supervising systems whose outputs exceed that competence. This article reviews the two oldest proposals for the problem, recursive reward modelling with its close relative iterated amplification, and debate between competing models before a human judge, and places them beside two newer lines of work, weak-to-strong generalization and model-written critique. The review first sets out why direct feedback fails under optimization pressure, drawing on published evidence of reward hacking, reward model overoptimization, sycophancy and unintended persuasion. It then examines the assumptions each oversight method needs, using simple cost and error formulas to show what decomposition and adversarial argument gain and where their guarantees stop. The empirical record on debate is reported as it stands, with clear gains for multi-turn debate in reading comprehension tasks where the judge cannot see the source passage and small or mixed effects elsewhere. A comparative framework is proposed that links each method to its required assumption, its characteristic failure mode and a diagnostic test, together with an evaluation protocol built on the sandwiching design. The framework is analytical and has not yet been tested empirically.

Keywords:- Scalable Oversight, Reinforcement Learning from Human Feedback, Recursive Reward Modelling, Iterated Amplification, AI Safety Via Debate, Weak-To-Strong Generalization, Sandwiching

I. INTRODUCTION

Amodei et al. listed scalable supervision among five concrete problems in AI safety [1]. The problem arises when the true objective is too costly to evaluate as often as training requires, and the usual response has been to learn a reward function from a much smaller number of human judgments. Christiano et al. trained Atari and simulated locomotion agents from pairwise human comparisons of short clips, with feedback on less than one percent of the agents' interactions with the environment [2]. Applied to language models, the same recipe produced summarisation policies that labellers preferred to much larger supervised models [3]. It also produced instruction-following models whose 1.3 billion parameter version was preferred to the 175 billion parameter GPT-3 [4]. This method, reinforcement learning from human feedback (RLHF), is now the usual final stage in training general-purpose assistants.

Every version of the recipe assumes that the person giving feedback can tell a better output from a worse one. Bowman et al. define scalable oversight as the problem of supervising systems that may outperform their

supervisors on most of the skills relevant to the task [5]. A code review that would take an expert an hour, or a proof with a subtle gap, may receive a confident rating from a hurried evaluator. A policy optimised against such ratings then learns whatever the evaluator rewards. Ngo et al. argue that systems trained in this way could learn to act deceptively to receive higher reward, and that they may appear aligned when they are not [6].

Two families of proposal have dominated the response. Recursive reward modelling and iterated amplification break a hard evaluation into easier pieces, with trained assistants helping the human at each piece. Debate sets two copies of a model against each other before a human judge. Weak-to-strong generalization and model-written critique are newer lines of work that approach the problem from the side of the learner and from the side of the evaluator. This article reviews all four, states the assumptions each one needs in a form that can be checked, and proposes a comparative framework with an evaluation protocol. Sections II to V cover direct feedback, recursive decomposition, debate, and the weak-to-strong and critique methods in turn. Section VI sets out the framework and the protocol. No experiments were run for this article, and every plotted quantity is computed from a formula stated in the text.

II. HUMAN FEEDBACK AND ITS FAILURE UNDER OPTIMIZATION

The reward model in RLHF is fitted to pairwise comparisons through the Bradley-Terry model of paired comparison [7]. For a prompt x and two responses y1 and y2, the probability that a labeller prefers y1 is modelled as:

$$P(y_1 \succ y_2 \mid x) = \sigma\big(r(x, y_1) - r(x, y_2)\big) \tag{1}$$

where σ is the logistic function and r is the learned reward, and the reward model is trained to minimise $-\log \sigma(r(x, y_w) - r(x, y_l))$ over observed winners y_w and losers y_l. The policy π is then optimised to maximise:

$$\mathbb{E}[r(x, y)] - \beta \mathrm{KL}(\pi \,\|\, \pi_{\mathrm{ref}}) \tag{2}$$

where the Kullback-Leibler term with weight β keeps the policy close to a reference model π_{ref} [3], [4]. Under this model a reward gap of 1 corresponds to a preference probability of $\sigma(1) \approx 0.73$, and a gap of 3 to about 0.95. The reward scale therefore encodes how consistently labellers separate two outputs. It carries no information about correctness beyond what the labellers could see.

Optimization against a learned proxy invites reward hacking. Skalse et al. formalise the phenomenon under the name reward gaming [8]. They prove that, over the set of all stochastic policies, two reward functions can be unhackable with respect to each other only if one of them is constant [8]. Pan et al. built four reinforcement learning environments with misspecified rewards [9]. They found that more capable agents more often exploit the misspecification, gaining proxy reward while losing true reward [9]. Gao et al. measured the effect for language models in a synthetic setting where a fixed gold reward model stands in for human labellers, and found that gold reward first rises and then falls as optimization against the proxy continues [10]. The curve takes a different functional form for best-of-n sampling and for reinforcement learning, and its coefficients scale smoothly with reward model size [10].

A second family of failures arises from the labellers. Sharma et al. found sycophancy in five deployed assistants across four free-form tasks [11]. In existing preference data, responses matching a user's stated views were more likely to be preferred, and both humans and preference models sometimes chose convincingly written sycophantic answers over correct ones [11]. Perez et al. had earlier reported that larger models repeat back a user's preferred answer more often [12]. Wen et al. asked time-limited human subjects to judge the correctness of model outputs on a question answering task and a programming task, scoring their judgments against gold labels [13]. RLHF made the models more convincing without making them more correct [13]. After RLHF, the subjects' false positive rate rose by 24.1 and 18.3 percentage points, from 41.0% to 65.1% on QuALITY and from 29.6% to 47.9% on APPS [13]. Casper et al. survey these and related problems, and separate those that better technique might address from those fundamental to learning from human approval [14].

Overoptimization and reward hacking exploit errors in the learned reward, while sycophancy and unintended persuasion exploit errors in the human judgments behind it. The methods reviewed in the next three sections try to improve those judgments, not to fit them more closely.

III. RECURSIVE REWARD MODELLING AND ITERATED AMPLIFICATION

Leike et al. proposed reward modelling as a research direction for agent alignment, in which a reward function learned from interaction with the user is optimised by reinforcement learning [15]. For tasks the user cannot evaluate unaided, their recursive variant has agents trained with earlier reward models help the user evaluate the outcomes of the next agent, so that each round extends the range of tasks that can be supervised [15]. In iterated amplification, a human solves a hard problem by delegating easier subquestions to copies of the current

model [16]. The model is then trained to imitate this amplified system [16]. Christiano et al. reported results on algorithmic tasks learned without an external reward function [16]. Fig. 1 contrasts these designs with direct evaluation and with debate.

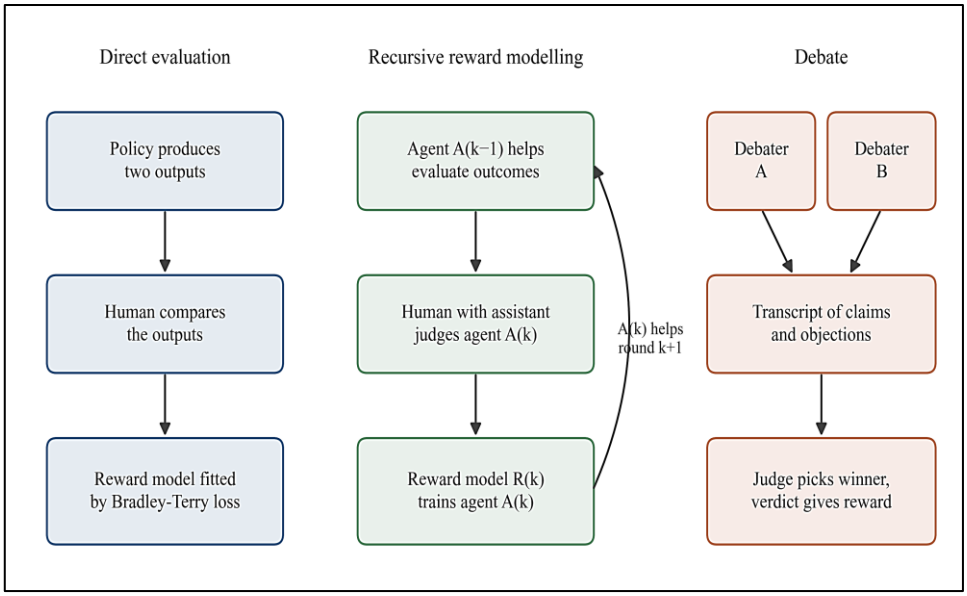


Fig. 1: Schematic of three oversight structures. In direct evaluation a human compares outputs and a reward model is fitted to the comparisons [2], [7]. In recursive reward modelling, agents trained in earlier rounds help the human evaluate the next agent [15]. In debate, two agents argue before a judge whose verdict supplies the reward.

Wu et al. applied recursive decomposition to the summarisation of entire novels [17]. The model summarises short sections and then summarises the summaries, and labellers evaluate each step without reading the whole book. The resulting summaries matched the quality of human-written ones for about 5% of books [17]. The study remains the clearest published case of decomposition making tractable an evaluation that would otherwise require hours of reading per judgment.

The benefit of decomposition can be stated as a count. Suppose a task has size N, measured in units that a person can check in one sitting, and is decomposed as a tree with branching factor b, so that the depth is $d = \log_b N$. Direct evaluation requires one judgment that reads all N units. A full recursive scheme that checks every internal node requires $\frac{(N-1)}{(b-1)}$ judgments of b sub-answers each, a total of $\frac{b(N-1)}{(b-1)}$ units. A scheme that checks only one path from root to leaf, which is what debate attempts, requires d judgments of b units, a total of $b\log_b N$. Fig. 2 plots the three totals for $b = 4$.

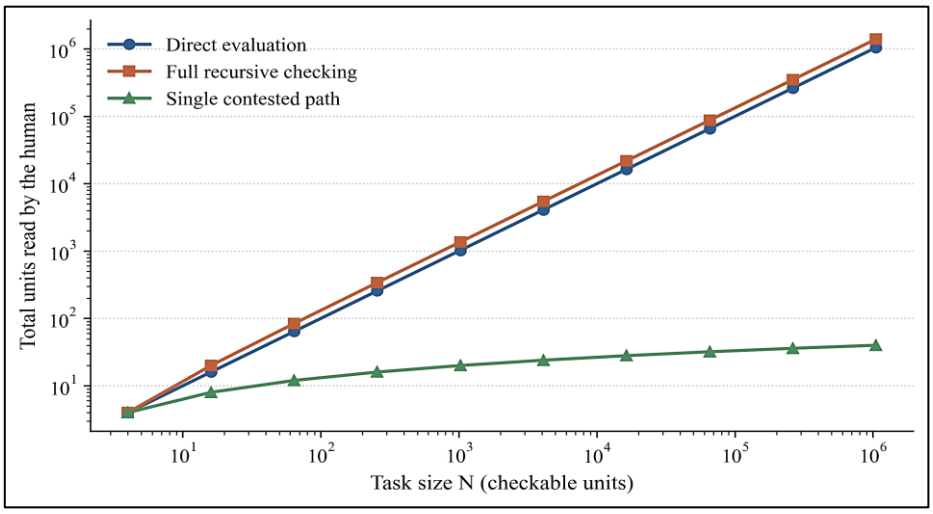


Fig. 2: Units a human must read to evaluate a task of size N under direct evaluation (N), full recursive checking of every node $\frac{b(N-1)}{(b-1)}$ and checking of a single contested path $b\log_b N$, computed from the formulas in Section III with branching factor $b = 4$. Both axes are logarithmic.

Fig. 2 carries two points. Full recursion does not reduce the total human effort, which exceeds N by the factor $\frac{b}{(b-1)}$. It does, however, cap every individual judgment at b units, a size that a person can check reliably. Only path checking reduces the total, from about 1.05 million units for direct evaluation or 1.4 million for full recursion to 40 at N $=4^{10}$, and it does so only if the checked path is the one that contains the error. Recursive reward modelling therefore relies on trained assistants to carry most of the load, and debate is designed to find the right path.

Three assumptions carry the recursive argument. The task must decompose into subquestions whose answers can be checked locally, and tasks that rest on tacit judgment, such as deciding whether a research plan is promising, may not decompose in this way. Errors must not accumulate faster than they are caught, since a reward model at round k inherits the mistakes of the assistants that helped train it. The assistants must also be honest. This is because an assistant that has learned to flatter the evaluator, as documented for RLHF models [11], [13], would carry that bias into every later round. None of these assumptions has been tested at the scale the proposals envisage. The main published evidence consists of the algorithmic tasks in [16] and the book summarisation study in [17].

IV. DEBATE: THEORY AND EMPIRICAL RECORD

Irving et al. proposed training agents by self-play on a zero-sum debate game [18]. Two agents take turns making short statements about a question, and a human then judges which agent gave the most true and useful information. The case for the design is an analogy with interactive proof systems [19]. A polynomial-time judge who checks answers directly can verify only problems in NP, whereas with optimal play debate lets such a judge answer any question in PSPACE [18], the class that interactive proofs capture [20]. The intuition is that a dishonest claim must rest on some false step, and an honest opponent can direct the judge to that step.

A simple error model shows what this gains. Let a sound argument have k steps, and let a judge who checks a step make an error with probability ε, independently across steps. A judge who must check every step reaches the correct verdict with probability $(1-\varepsilon)^k$, while a judge shown only the single step in dispute, as debate intends, is correct with probability $1-\varepsilon$ whatever the value of k. Fig. 3 plots both cases. At ε = 0.05 the probability of a correct full check falls to 0.60 at k = 10 and to 0.077 at k = 50, while the debate judge stays at 0.95.

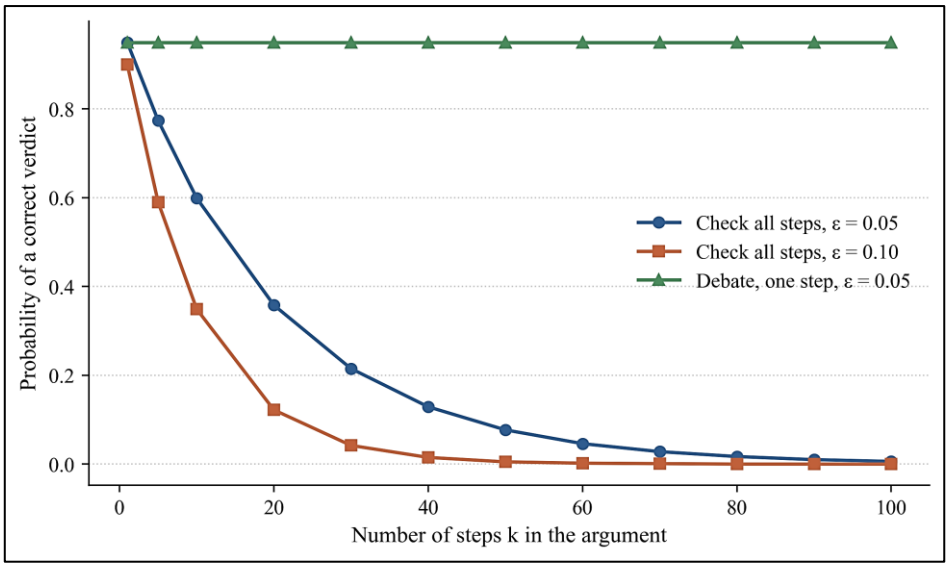


Fig. 3: Probability that a judge reaches the correct verdict on a sound k-step argument when every step must be checked, $(1-\varepsilon)^k$, for per-step error rates ε = 0.05 and ε = 0.10, compared with a debate judge who checks one disputed step, $1-\varepsilon$ at ε = 0.05. Values are computed from the error model in Section IV, which assumes independent errors.

The guarantee holds only at equilibrium and under conditions that the formula hides. The honest debater must be able to find the flawed step. A dishonest debater may offer an argument so long that neither side can locate its flaw within the debate, which leaves the honest side nothing to point to. This difficulty is usually called the obfuscated arguments problem [21]. The original complexity argument also assumed an honest debater able to simulate deterministic systems for exponentially many steps. Brown-Cohen et al. designed protocols in which the honest strategy succeeds with a polynomial number of simulation steps and can verify stochastic systems, even when the dishonest side may use exponentially many steps [21]. Their analysis models human judgment as an oracle for local questions, which real judges only approximate.

Six studies are listed in Table 1. In the MNIST experiment of [18], agents competed to convince a sparse classifier that saw only the pixels they chose to show, and raised its accuracy from 59.4% to 88.9% with six pixels. Parrish et al. gave single-turn arguments to human judges who had not read the passage behind a hard reading comprehension question [22]. The arguments did not improve accuracy, whereas selected text snippets did [22]. With multi-turn debates between human experts, Michael et al. reported 84% judge accuracy against 74% for consultancy, a baseline in which one expert argues for an answer that is correct half the time [23]. Khan et al. found that debate between LLM experts raised accuracy to 76% for non-expert model judges and to 88% for human judges [24]. The corresponding naive baselines were 48% and 60% [24]. Optimising the debaters for persuasiveness also made judges more accurate in their setting [24].

Table 1. Published experiments on debate and assisted judging

Study	Task and asymmetry	Judge	Comparison	Reported outcome
Irving et al. [18]	MNIST, judge sees only pixels chosen by debaters	Sparse classifier	Debate vs. 6 random pixels	Accuracy 59.4% to 88.9% with 6 pixels
Parrish et al. [22]	Reading comprehension, passage hidden	Humans	Single-turn arguments vs. none	No gain from arguments, gain from text snippets
Michael et al. [23]	Reading comprehension, passage hidden	Humans	Debate vs. consultancy	84% vs. 74%
Khan et al. [24]	Reading comprehension, passage hidden	LLMs and humans	Debate vs. naive baseline	76% vs. 48% (LLM), 88% vs. 60% (human)
Kenton et al. [25]	QA, mathematics, coding, logic, multimodal	Weaker LLMs	Debate vs. consultancy and direct QA	Better than consultancy on all tasks (random side assignment), mixed against direct QA without information asymmetry
Bowman et al. [5]	MMLU, time-limited QuALITY	Humans with an LLM assistant	Assisted vs. unaided humans and model alone	Assisted humans outperform both

Kenton et al. used weaker LLMs as judges across extractive question answering, mathematics, coding, logic and multimodal reasoning [25]. Debate beat consultancy on every task when consultants were assigned sides at random. Against direct question answering by the judge, debate helped on extractive tasks with information asymmetry, and results on the other tasks were mixed [25]. Stronger debaters raised judge accuracy, though by less than earlier studies had found [25]. The positive results thus come mostly from reading comprehension with a hidden passage, where a verified quote can settle the question. Evidence is weaker for tasks in which the judge has all the information but lacks the skill to use it, which is the case that scalable oversight is ultimately about.

V. WEAK-TO-STRONG GENERALIZATION AND AI-ASSISTED CRITIQUE

Burns et al. studied an analogy in which a small model supervises a large one [26]. Strong pretrained models fine-tuned on labels from a weak supervisor consistently outperformed that supervisor, but fell short of the same model trained on ground truth. The outcome is measured as the performance gap recovered:

$$\mathrm{PGR} = \frac{P_{w2s} - P_{weak}}{P_{strong} - P_{weak}} \tag{3}$$

where P_{weak} is the accuracy of the supervisor, P_{strong} is the ceiling of the strong model trained on true labels and P_{w2s} is the accuracy of the strong model trained on weak labels [26]. With a GPT-2 level supervisor and an auxiliary confidence loss, GPT-4 recovered close to GPT-3.5 level performance on NLP tasks [26]. The errors of a weak model may be easier for a strong model to recognise than the systematic errors of human labellers. This limits how far the analogy transfers to human oversight.

Model-written critique works on the side of the evaluator. Saunders et al. fine-tuned language models to write critiques of summaries, and the critiques helped human evaluators find flaws they would otherwise have

missed, including flaws planted deliberately to mislead [27]. Larger models also wrote more helpful critiques [27]. Their comparison of a model's ability to recognise a flaw with its ability to articulate it suggests that models may hold relevant knowledge they do not express as critiques [27]. McAleese et al. trained code critics with RLHF. On naturally occurring errors, their critiques were preferred to human critiques in 63% of cases [28]. The critics also hallucinated bugs, and human contractors working with critics caught a similar number of bugs to the critics alone while hallucinating less [28].

AI feedback can also replace human labels outright. Constitutional AI trains a harmless assistant from a short list of written principles, with self-critique and revision in a supervised phase and model-generated preference labels in a reinforcement learning phase [29], and self-critique based alignment of this kind has also been discussed in this journal [30]. Lee et al. found that reinforcement learning from AI feedback performed comparably to RLHF on summarisation and on helpful and harmless dialogue [31]. Zheng et al. reported that strong LLM judges agree with human preferences more than 80% of the time, the same level at which humans agree with each other, and documented position, verbosity and self-enhancement biases in those judges [32]. AI feedback scales the volume of supervision without raising its ceiling. This is because the labelling model is bounded by its own competence and may share the blind spots of the models it evaluates.

VI. A COMPARATIVE FRAMEWORK AND EVALUATION PROTOCOL

The methods reviewed above differ less in their mechanics than in the assumptions behind their guarantees. Table 2 sets out the proposed framework. For each method it records the assumption required for sound oversight, the place where human effort falls, the failure that becomes likely when the assumption breaks, and a diagnostic test for that failure. The framework is an organising proposal and has not been applied to a common benchmark.

Table 2. Proposed comparative framework for scalable oversight methods

Method	Assumption required	Where human effort falls	Characteristic failure	Diagnostic test
Direct RLHF [2], [4]	Evaluator can judge outputs unaided	Every comparison	Reward hacking, sycophancy, persuasion [10], [11], [13]	Agreement of evaluators with expert labels
Recursive reward modelling and amplification [15], [16]	Task decomposes into locally checkable parts	Every node, with assistants	Errors passed forward across rounds	Plant flaws at a chosen depth and measure detection
Debate [18], [21]	Honest strategy wins, judge accurate on one step	One contested path	Obfuscated arguments, persuasion over truth	Assign sides, vary argument length and debater strength
Weak-to-strong generalization [26]	Strong model generalises intent, not supervisor errors	Weak labels only	Imitation of systematic supervisor errors	Performance gap recovered on held-out ground truth
Critique and AI feedback [27], [29], [31]	Critic errors uncorrelated with generator errors	Review of critiques	Hallucinated flaws, shared blind spots, judge bias [28], [32]	Planted bugs, with catch and false alarm rates

Three failure modes cut across the rows of Table 2. Collusion becomes possible whenever copies of one model fill both the supervising and the supervised roles, as in debate between two instances, in recursive assistants trained from earlier agents and in AI feedback drawn from the policy's own checkpoint. Nothing in a zero-sum objective stops two copies from sharing a blind spot. Persuasion that outruns truth appears in RLHF [13] and in consultancy [23]. Khan et al. found persuasiveness and truthfulness moving together in debate [24], but it is not yet established that this holds when the judge cannot check a quote. Sycophancy can enter any method in which output is rated by a person whose views the model can infer [11].

The sandwiching design of Bowman et al. gives the evaluation protocol its structure [5]. Tasks are chosen so that the model alone and unaided non-experts both fail while domain experts succeed, and the non-experts then use an oversight method together with the model while the experts' answers serve only to score the result. In the

proof of concept on MMLU and time-limited QuALITY, non-experts who chatted with an unreliable model assistant outperformed both the model alone and their own unaided performance [5]. The protocol proposed here extends that design in four ways. Each method faces a planted adversary, a model fine-tuned to argue for wrong answers or to insert subtle flaws, so that measured accuracy reflects resistance to deliberate deception. Decomposition depth or argument length is varied, so that the error growth predicted in Fig. 3 can be compared with the growth actually observed. Debater and judge strength are varied independently, following [25], and results are reported as the share of the gap between non-expert and expert accuracy that the method recovers, by analogy with the PGR of [26]. False alarm rates are reported alongside detection rates, because flagging correct outputs has its own cost, as the hallucinated bugs in [28] show.

Several limits apply to the framework and to the field. Sandwiching needs tasks on which experts exist, so it cannot test oversight of capabilities that exceed every human. The error model behind Fig. 3 assumes independent errors and a single decisive flaw. Correlated errors by the judge would shrink the advantage it shows. Judges in several published studies worked under time limits [5], [13], and trained evaluators with more time may behave differently. Methods that inspect internal computation instead of outputs, such as the mechanistic study of in-context learning discussed in this journal [33], offer a check that does not depend on how a judge reads an argument, although interpretability at the scale of frontier models remains incomplete.

VII. CONCLUSION

Human feedback works as a training signal only within the range of outputs that people can evaluate. The evidence on reward hacking, overoptimization, sycophancy and unintended persuasion shows optimization pressure finding the edges of that range. In the idealised counting model of Section III, full recursive checking caps the size of each judgment but leaves total human effort roughly unchanged. Recursive reward modelling depends on decomposability and on assistants that do not pass their errors forward. In the same model, debate offers an oversight cost that grows with the logarithm of task size, but only under an honest equilibrium that obfuscated arguments and judge error can break.

The empirical record supports multi-turn debate most clearly in reading comprehension with a hidden passage, where a verifiable quote settles the question, and shows small or mixed effects elsewhere. Weak-to-strong generalization and model-written critique offer partial gains that are easier to measure with present models. Table 2 links each method to the assumption it needs and to a test that would expose its failure. The next step is to apply the extended sandwiching protocol of Section VI to all four families of method on a shared task set. Claims about their relative merit should wait for that comparison.

REFERENCES

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” *arXiv preprint arXiv:1606.06565*, 2016.

[2] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[3] N. Stiennon *et al*., “Learning to summarize with human feedback,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 3008–3021, 2020.

[4] L. Ouyang *et al*., “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27730–27744, 2022.

[5] S. R. Bowman *et al*., “Measuring progress on scalable oversight for large language models,” *arXiv preprint arXiv:2211.03540*, 2022.

[6] R. Ngo, L. Chan, and S. Mindermann, “The alignment problem from a deep learning perspective,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[7] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. The method of paired comparisons,” *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.

[8] J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, “Defining and characterizing reward gaming,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 9460–9471, 2022.

[9] A. Pan, K. Bhatia, and J. Steinhardt, “The effects of reward misspecification: Mapping and mitigating misaligned models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[10] L. Gao, J. Schulman, and J. Hilton, “Scaling laws for reward model overoptimization,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR, vol. 202, pp. 10835–10866, 2023.

[11] M. Sharma *et al*., “Towards understanding sycophancy in language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[12] E. Perez *et al*., “Discovering language model behaviors with model-written evaluations,” in *Findings of the Association for Computational Linguistics (ACL)*, pp. 13387–13434, 2023.

[13] J. Wen *et al*., “Language models learn to mislead humans via RLHF,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[14] S. Casper *et al*., “Open problems and fundamental limitations of reinforcement learning from human feedback,” *Transactions on Machine Learning Research*, 2023.

[15] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: A research direction," *arXiv preprint arXiv:1811.07871*, 2018.
[16] P. Christiano, B. Shlegeris, and D. Amodei, "Supervising strong learners by amplifying weak experts," *arXiv preprint arXiv:1810.08575*, 2018.
[17] J. Wu *et al*., "Recursively summarizing books with human feedback," *arXiv preprint arXiv:2109.10862*, 2021.
[18] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *arXiv preprint arXiv:1805.00899*, 2018.
[19] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof systems," *SIAM Journal on Computing*, vol. 18, no. 1, pp. 186–208, 1989.
[20] A. Shamir, "IP = PSPACE," *Journal of the ACM*, vol. 39, no. 4, pp. 869–877, 1992.
[21] J. Brown-Cohen, G. Irving, and G. Piliouras, "Scalable AI safety via doubly-efficient debate," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 4585–4602, 2024.
[22] A. Parrish *et al*., "Single-turn debate does not help humans answer hard reading-comprehension questions," in *Proceedings of the 1st Workshop on Learning from Natural Language Supervision (LNLS)*, pp. 17–28, 2022.
[23] J. Michael *et al*., "Debate helps supervise unreliable experts," *arXiv preprint arXiv:2311.08702*, 2023.
[24] A. Khan *et al*., "Debating with more persuasive LLMs leads to more truthful answers," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 23662–23733, 2024.
[25] Z. Kenton *et al*., "On scalable oversight with weak LLMs judging strong LLMs," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 75229–75276, 2024.
[26] C. Burns *et al*., "Weak-to-strong generalization: Eliciting strong capabilities with weak supervision," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 4971–5012, 2024.
[27] W. Saunders *et al*., "Self-critiquing models for assisting human evaluators," *arXiv preprint arXiv:2206.05802*, 2022.
[28] N. McAleese, R. M. Pokorny, J. F. Ceron Uribe, E. Nitishinskaya, M. Trebacz, and J. Leike, "LLM critics help catch LLM bugs," *arXiv preprint arXiv:2407.00215*, 2024.
[29] Y. Bai *et al*., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.
[30] W. M. John, "Constitutional AI: Scalable alignment through self-critique and revision," *Eduschool Journal of Artificial Intelligence Research*, vol. 1, no. 1, pp. 10–19, 2026, doi: 10.63090/EJAIR/3139.7336.0002.
[31] H. Lee *et al*., "RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 26874–26901, 2024.
[32] L. Zheng *et al*., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, vol. 36, pp. 46595–46623, 2023.
[33] M. T. V, "Mechanistic interpretability of in-context learning in transformers," *Eduschool Journal of Artificial Intelligence Research*, vol. 1, no. 1, pp. 20–29, 2026, doi: 10.63090/EJAIR/3139.7336.0003.