ORIGINAL RESEARCH **Open access**

A Theory of Justified Abstention for Uncertain Materials AI Models

Omar Khalid[1*], Sara Nadeem[1], Bilal Farooq[2], Hina Saeed[1]

Abstract

Materials AI models invariably produce predictions for every input, even when operating under high uncertainty, distributional shifts, or conditions where the potential costs of error far outweigh any informational benefit. This reflexive prediction habit represents a critical gap in the field, as models rarely—if ever—choose to abstain despite the high-stakes nature of materials discovery, where erroneous outputs can trigger wasteful synthesis campaigns, compromise safety assessments for novel compounds, or mislead decisions involving rare-event phenomena such as phase instabilities under extreme conditions. Justified abstention is defined here as the deliberate, epistemically grounded decision by a model to withhold any prediction when the expected utility of outputting a value falls below the utility of remaining silent, thereby prioritizing scientific integrity over forced coverage. This paper articulates a novel theory of justified abstention built on three core principles—competence boundary, risk threshold, and resource consideration—alongside five explicit operational criteria that together provide a principled framework for when abstention becomes not only permissible but obligatory in materials contexts. Four distinct types of abstention are delineated (input-based, prediction-based, risk-based, and resource-based), each with clear triggers and materials-specific illustrations that underscore their necessity. The implications extend to transformed design pipelines, where abstention mechanisms foster greater trustworthiness, enable more efficient allocation of experimental resources, and shift materials AI from indiscriminate oracles to responsible scientific partners capable of signaling their own epistemic limits. By embedding justified abstention as a core design feature rather than an afterthought, the framework addresses a longstanding oversight in the literature. It offers a pathway toward more reliable, ethically defensible AI systems for materials science.

Keywords Materials AI, Uncertainty quantification, Justified abstention, Selective prediction, Out-of-distribution detection, Risk-aware modeling

*Correspondence:
Omar Khalid
omar.khalid@gmail.com
[1] Department of Intelligent Materials Analytics, King Fahd University, Dhahran, Saudi Arabia
[2] Department of AI Engineering Materials, Qatar University, Doha, Qatar

Introduction

Materials AI models have become indispensable tools across computational materials science, generating property predictions, stability assessments, and synthesis recommendations with remarkable speed and apparent precision [1]. Yet a fundamental and largely unexamined assumption underlies nearly all such systems: every input must receive a prediction, regardless of the model's internal confidence, the distance of the query from its training distribution, or the downstream consequences of potential error. This paper identifies the absence of justified abstention as a central gap in materials AI. Models rarely refuse to make predictions even when operating in regimes of high uncertainty or out-of-distribution conditions. This practice stands in contrast to broader machine learning research on selective classification and reject options [2-7]. The consequences are particularly acute in materials science, where a single erroneous prediction can cascade into costly experimental validation failures, safety oversights in handling unstable or toxic compounds, or

missed opportunities in exploring rare but transformative material phenomena [8-15].

The problem is not merely technical but epistemological. Materials discovery operates under conditions of inherent sparsity and high experimental cost; training data are expensive to acquire, and many critical regimes lie far from dense regions of available data [1]. When models nonetheless emit outputs without acknowledging epistemic boundaries, they implicitly claim a level of competence they do not possess [2]. Selective classification frameworks have demonstrated that allowing models to abstain can improve reliability without sacrificing coverage in supported regions [3, 5, 11]. Yet materials AI has largely ignored this possibility, defaulting instead to full-coverage prediction pipelines [1].

The Abstention Problem

The abstention problem in materials AI arises from the mismatch between the epistemic limits of any finite model and the practical demands of high-stakes materials decision-making. Materials discovery routinely involves inputs that lie outside the support of training distributions—whether because a candidate composition is chemically novel, a processing condition is extreme, or a target property has few experimental analogs [1, 12, 13]. In such cases, forcing a prediction risks propagating errors that are not merely inaccurate but actively misleading [14, 16, 17]. Costly experiments are a primary concern: an erroneous stability prediction may trigger months of unsuccessful synthesis attempts, consuming both time and scarce reagents. Safety-critical applications compound the issue; predictions about the reactivity or toxicity of new intermetallics carry direct implications for handling protocols and regulatory approval [15, 18-22]. When models lack relevant training data for rare events, continued prediction becomes epistemically irresponsible [2].

Distributional shift further exacerbates the problem. Materials AI systems trained on curated databases frequently encounter queries drawn from entirely different regions of chemical or structural space. Yet, current pipelines offer no mechanism to decline such inputs gracefully [1, 12, 13]. The default behavior—always producing an output—implicitly assumes that coverage is an unconditional good. This assumption fails under scrutiny. In scientific decision-making, the utility of a prediction must account for asymmetric costs of error in real-world deployment [7, 19]. Predicting a metastable phase as stable, for example, may lead to experimental dead-ends far more expensive than abstaining.

The deeper issue is normative. Materials science values cautious inference and transparent acknowledgment of uncertainty [2]. A model that never abstains violates this norm by presenting every output as equally trustworthy. Abstention serves as an epistemic signal that the model has reached the boundary of its justified domain and invites human intervention or further data collection [18, 22]. Strategic abstention can therefore increase the overall trustworthiness of AI-assisted discovery pipelines [3, 8].

Current Practices

Current practices in materials AI overwhelmingly favor unconditional prediction. Even when uncertainty quantification techniques are employed, models still emit a final value rather than withholding judgment [1, 14]. Out-of-distribution detection methods have begun to appear, flagging inputs that deviate from training support, yet these flags rarely translate into systematic abstention; the model typically proceeds to predict anyway [12, 13, 17]. Confidence thresholds exist in isolated implementations but remain ad hoc and lack principled grounding [8, 10].

For instance, structure-based out-of-distribution benchmarks acknowledge the challenge yet continue to prioritize full coverage over selective refusal [1]. Similarly, density-based detection approaches identify anomalies but stop short of recommending abstention as a design feature [12]. In broader machine learning, selective classification has matured into a coherent subfield with explicit reject-option frameworks that balance accuracy against coverage [3, 5-7]. Materials AI has not adopted these advances in a principled manner.

Uncertainty quantification is typically treated as a diagnostic rather than a decision mechanism [14, 18-22]. As a result, models operate beyond their competence regions while still producing outputs. Philosophical and ethical discussions of abstention remain largely disconnected from material applications [2], leaving a structural gap: the field has tools for detecting uncertainty and distributional shift. Still, it lacks a framework to act on them.

A Theory of Justified Abstention

This paper proposes a theory of justified abstention specifically calibrated to the epistemic and practical realities of materials AI. The theory rests on three interlocking principles that together define when withholding a prediction is the rational and responsible choice.

Competence boundary

A materials AI model possesses a bounded region of competence defined by the joint distribution of its training data and the inductive biases encoded in its architecture. Outside this region—whether due to compositional novelty, extreme processing parameters, or underrepresented physical phenomena—the model's outputs lack epistemic warrant. Abstention is therefore justified whenever an input violates the competence boundary, preserving the integrity of scientific inference by refusing to extrapolate beyond justified domains.

Risk threshold

Even within the competence boundary, certain predictions carry asymmetric risks whose expected harm exceeds acceptable scientific or societal tolerances. The theory, therefore, requires models to evaluate whether the potential negative consequences of an incorrect prediction (in terms of experimental cost, safety, or downstream decision error) surpass a predefined risk threshold. When this occurs, abstention becomes obligatory to prevent the propagation of high-consequence misinformation.

Resource consideration

Materials discovery is resource-constrained; laboratory validation, high-performance computing cycles, and human expert time are all finite. The theory incorporates an economic dimension: abstention is justified when the expected resource expenditure required to act upon a low-utility prediction outweighs the informational benefit provided. This principle ensures that AI systems operate sustainably within the broader ecosystem of scientific research.

Justified abstention

The deliberate withholding of a prediction when the expected utility of predicting is less than the expected utility of abstaining. Expected utility here integrates predictive confidence, consequence severity, and resource costs within the materials science context.

Figure 1 presents the hierarchical decision architecture of justified abstention, showing how foundational principles are translated into sequential abstention checks that culminate in either scientifically warranted prediction or justified refusal.

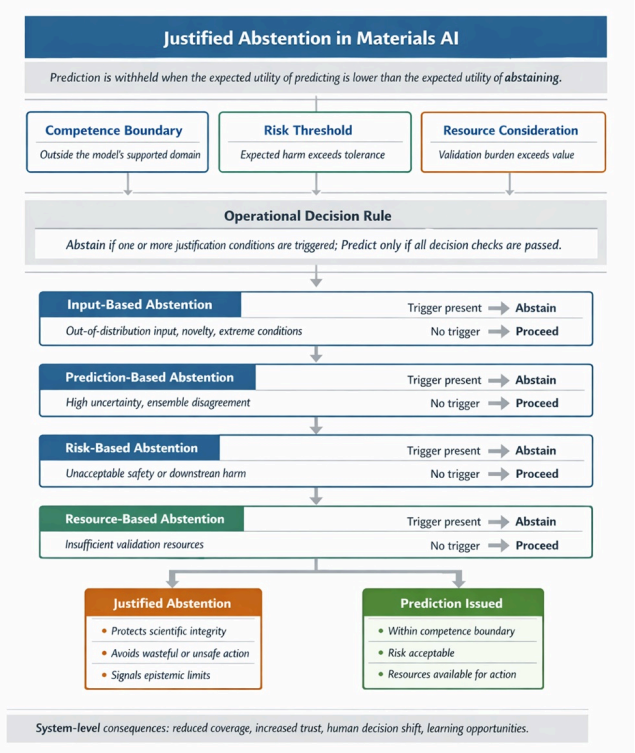

Figure 1. The hierarchical decision architecture of justified abstention.

The abstention decision process can be conceptualized as a sequential decision flow: an input material representation first encounters the competence boundary check (principle 1); if it lies outside the supported region, the model abstains immediately. If the input passes this gate, the risk threshold (principle 2) is evaluated against domain-specific harm criteria; exceeding the threshold again triggers abstention. Finally, resource consideration (principle 3) weighs the projected validation burden; only inputs that survive all three checks receive a prediction.

Table 1 consolidates the analytical architecture of justified abstention by linking the theory's normative foundations to its operational criteria and ultimate decision consequences.

Table 1. Analytical architecture of justified abstention in materials AI: principles, operational criteria, and decision function

Analytical layer	Core element	Definition in this manuscript	Primary evaluative question
Normative foundation	Competence boundary	A model is justified only within the domain supported by its training distribution and inductive biases	Is this input within the model's epistemically warranted domain?
Normative foundation	Risk threshold	A model should not predict when the expected harm of error exceeds an acceptable tolerance	Could a wrong prediction generate unacceptable scientific, safety, or regulatory harm?
Normative foundation	Resource consideration	A model should not issue low-utility outputs that cannot be responsibly acted upon	Do available resources justify acting on this prediction?
Formal decision rule	Justified abstention	Prediction is withheld when the expected utility of predicting is lower than the expected utility of abstaining	Is prediction less valuable than silence in this case?
Operational criterion	Uncertainty threshold	High predictive uncertainty makes the output scientifically unreliable	Is uncertainty above calibrated tolerance?
Operational criterion	Out-of-distribution score	Inputs too distant from training support lack epistemic warrant	Is the query sufficiently anomalous relative to training data?
Operational criterion	Cost asymmetry	False predictions may be more damaging than correct predictions are beneficial	Is the cost of error asymmetrically large?
Operational criterion	Consequence severity	Some predictions are too high-stakes to issue under uncertainty	Would an incorrect output create severe downstream harm?
Operational criterion	Resource availability	A prediction without a feasible follow-up may impose an unjustified burden	Can this output realistically be validated or acted upon?
System output	Prediction issued	Output is released only when all relevant checks are passed	Has the case survived competence, risk, and resource evaluation?
System output	Justified abstention	Output is withheld as a	Has at least one decisive abstention

		scientifically responsible act	condition been triggered?

This conceptual flowchart ensures that abstention emerges as a reasoned outcome rather than an arbitrary cutoff, aligning model behavior with the normative demands of scientific reliability.

Criteria for Justified Abstention

To operationalize the theory, five explicit criteria are articulated. Each criterion functions as a testable condition that, when satisfied, renders abstention justified.

Uncertainty threshold

Abstention is required whenever predictive uncertainty—quantified through ensemble variance, Bayesian posterior width, or conformal scores—exceeds a calibrated threshold derived from validation on in-distribution held-out data. This criterion directly links epistemic humility to model behavior, ensuring that outputs are issued only when the model can quantify its confidence at scientifically acceptable levels.

Out-of-distribution score

Inputs yielding an out-of-distribution score above a predefined cutoff, as measured by density estimation, reconstruction error, or embedding distance metrics, trigger abstention. This criterion enforces the competence boundary at the input level, preventing the model from issuing predictions on chemically or structurally alien queries.

Cost asymmetry

Abstention is justified when the estimated cost of a false prediction (experimental follow-up, safety mitigation, or opportunity loss) exceeds the benefit of a correct prediction by a domain-specific factor, typically determined through stakeholder-elicited utility functions. This criterion incorporates the asymmetric economics inherent to materials research.

Consequence severity

Predictions whose potential downstream harm—measured against safety, environmental, or regulatory thresholds—surpass an acceptability threshold must be withheld. In materials contexts, this might include predictions that could lead to unstable compound recommendations or environmentally persistent materials without adequate mitigation data.

Resource availability

When downstream validation resources (computational, experimental, or human) fall below a minimum required level for meaningful follow-up, the model abstains. This criterion ensures that AI outputs remain actionable within realistic laboratory constraints rather than generating unverified claims that burden the broader research ecosystem.

These criteria are not isolated checkboxes but interdependent components of the overall theory. In practice, they are evaluated sequentially or in a weighted combination, with the formal definition of justified abstention serving as the integrative decision rule. By embedding these criteria directly into model architectures or inference pipelines, materials AI can move from reflexive prediction to principled, context-sensitive refusal.

Types of Abstention

The theory of justified abstention introduces a clear typology of four distinct types, each tailored to the epistemic and practical demands of materials AI. These types are not mutually exclusive but can be combined within a single decision pipeline, allowing models to evaluate inputs along multiple dimensions before deciding whether to predict or abstain.

Table 2 clarifies how the four abstention types differ in trigger logic, epistemic justification, and practical consequences, thereby translating the theory into a usable design typology for materials AI systems.

Table 2. Comparative typology of abstention modes in materials AI: trigger logic, scientific rationale, trade-offs, and pipeline implications

Abstention type	Immediate trigger	Primary epistemic rationale	Represent materials scenari
Input-based	Input falls outside the	The model lacks	Novel hig entropy a

abstention	supported training distribution	competence for chemically or structurally alien queries	quinary nit or extrer pressur regime ab from train data
Prediction-based abstention	Confidence falls below the calibrated threshold despite nominal in-distribution status	The model recognizes insufficient internal reliability for the specific output	Battery electroly conducti estimate w very wic uncertai interval strong ensemb disagreen
Risk-based abstention	Potential downstream harm exceeds the acceptable threshold	Scientific responsibility requires refusal in high-consequence settings	Therma stabilit prediction f aerospace where fail could ha major saf implicatic
Resource-based abstention	Validation burden exceeds available resources or expected value	Predictions should remain actionable within real laboratory and computational constraints	Promisir photovolt candidate pursue because validatic queue a resources already saturate
Cross-type interaction	Multiple triggers activate simultaneously	Abstention is often overdetermined rather than caused by a single factor	Novel composit with hig uncertair high valida cost, ar nontrivial s implicatic
System-level implication	Abstention becomes part of inference	Refusal is a deliberate scientific	The mate screenir pipelin

	design rather than a post hoc warning	action, not a model defect	explicitly rc abstaine cases to hu review

By distinguishing these categories, the framework provides designers with precise levers for embedding refusal mechanisms without resorting to arbitrary heuristics.

Input-based abstention is defined as the refusal to generate any prediction when the input material representation falls outside the supported region of the training distribution. The trigger is an elevated out-of-distribution score obtained through density estimation, reconstruction error in autoencoders, or embedding-space distance metrics [1, 12, 14, 17]. In materials contexts, this type becomes relevant when a query involves a high-entropy alloy composition never seen during training or an extreme processing condition such as gigapascal pressures far removed from database precedents. For instance, a model trained predominantly on binary and ternary oxides might abstain when presented with a quinary nitride under cryogenic conditions, thereby preventing spurious extrapolations that could mislead synthesis routes. The primary trade-off lies in reduced coverage: while the model declines a subset of potentially valuable queries, it gains substantial epistemic reliability and avoids the propagation of chemically implausible outputs. Materials scientists benefit because abstention here functions as an early warning, prompting targeted data collection rather than blind experimentation.

Prediction-based abstention occurs when the model's internal confidence in its own output falls below a calibrated threshold, even if the input appears in-distribution. Selective classification frameworks provide the foundational machinery for this type, wherein the model computes a confidence score—often via ensemble disagreement or posterior probability—and abstains if the score indicates insufficient reliability [3, 5, 8, 11]. In practice, this manifests in materials property prediction tasks where a model might produce a formation energy estimate but recognize that multiple plausible crystal structures yield conflicting values. A concrete materials example arises in battery electrolyte screening: the model predicts ionic conductivity for a candidate solvent but abstains when the uncertainty interval spans orders of magnitude, signaling that the prediction cannot reliably guide experimental prioritization.

The trade-off involves balancing coverage against accuracy. At the same time, some in-distribution queries remain unanswered, the predictions that are issued exhibit markedly higher trustworthiness, aligning model behavior with scientific standards of reproducibility.

Risk-based abstention is activated when the anticipated downstream harm from an incorrect prediction exceeds a domain-specific risk threshold, irrespective of confidence or distributional status. This type draws on ethical and philosophical considerations of machine learning abstention, emphasizing that certain material decisions carry asymmetric consequences for safety or environmental impact [2, 19, 22]. Triggers include stakeholder-defined harm functions—such as toxicity scores or reactivity indices—that quantify potential negative outcomes. Consider a model evaluating a novel intermetallic for aerospace applications: even with moderate confidence in its thermal stability prediction, the system abstains if the risk of undetected phase decomposition under operational heat could lead to structural failure. The trade-off here is explicit: coverage is deliberately sacrificed in high-stakes regimes to protect human safety and regulatory compliance, yet this sacrifice enhances the model's role as a responsible partner rather than an unchecked oracle.

Resource-based abstention addresses the practical constraints of materials research by refusing prediction when the expected cost of downstream validation exceeds the informational benefit. The trigger is an assessment of available computational cycles, laboratory throughput, or human expert time, integrated into the utility calculation [20, 23]. In high-throughput virtual screening for photovoltaics, for example, the model might abstain from a promising perovskite candidate if experimental validation resources are already saturated with higher-priority leads, thereby preventing the generation of unactionable suggestions. This type acknowledges that materials AI operates within a finite ecosystem, producing outputs that cannot be followed up, waste collective resources, and erodes trust. The trade-off is temporal: short-term coverage decreases, but long-term efficiency improves as the system focuses effort where it can meaningfully advance discovery. Collectively, these four types transform abstention from a blunt instrument into a nuanced, context-sensitive capability that respects the unique epistemology of materials science.

Consequences of Abstention

Implementing justified abstention carries four primary consequences that reshape the materials AI pipeline at both technical and organizational levels. These consequences are not unintended side effects but deliberate design outcomes that enhance scientific integrity.

Reduced coverage manifests as a deliberate contraction in the fraction of inputs receiving predictions. While this may initially appear limiting, the reduction is epistemically beneficial; models cease to populate chemical space with low-utility extrapolations, thereby concentrating experimental resources on higher-confidence regions. In materials discovery workflows, this means fewer false leads are pursued, conserving reagents and instrument time.

Increased trust arises because every issued prediction has survived rigorous justification checks. Users—whether computational chemists or experimentalists—can rely on outputs with greater assurance, knowing that abstention has filtered out unreliable cases. This trust is particularly valuable in collaborative settings where AI suggestions inform grant proposals or patent filings; a model that abstains transparently earns credibility precisely by demonstrating self-awareness.

Decision shift requires human experts to engage more actively with abstained cases. Rather than passively accepting AI outputs, researchers must interpret abstention signals as invitations for additional data acquisition, literature review, or alternative modeling strategies. This shift elevates the role of domain expertise and prevents over-reliance on automation.

Learning opportunities emerge because abstention logs serve as diagnostic data for iterative model improvement. Patterns in abstained inputs highlight underrepresented regions of materials space, guiding targeted data collection campaigns or architectural refinements. Over time, the frequency and nature of abstentions become a measurable indicator of progress toward broader competence boundaries.

Taken together, these consequences reposition abstention as a generative force rather than a limitation, fostering more sustainable and trustworthy materials AI ecosystems.

Relation to Existing Concepts

Justified abstention builds upon but remains distinct from several established concepts in machine learning and

materials AI. It is not merely uncertainty quantification, which, while essential, stops at reporting confidence without enforcing refusal [14, 24-29]. Uncertainty quantification provides the probabilistic substrate—such as ensemble variance or Bayesian credible intervals—yet leaves the final decision to predict in the hands of the user. Justified abstention, by contrast, converts quantified uncertainty into an operational refusal when predefined criteria are met.

The framework also extends out-of-distribution detection without being reducible to it. Out-of-distribution methods excel at flagging anomalous inputs, yet most implementations in materials contexts continue to generate predictions alongside the flag [1, 12, 14, 17]. Justified abstention incorporates out-of-distribution scores as one criterion within a broader decision process, ensuring that detection leads to principled withholding rather than supplementary annotation.

Finally, justified abstention refines selective prediction and reject-option classification by grounding them in materials-specific utility functions. General selective classification frameworks balance accuracy and coverage through confidence thresholds [3, 5, 7, 10], but they rarely incorporate consequence severity or resource constraints. The present theory, therefore, subsumes selective prediction as a foundational mechanism while augmenting it with risk and resource principles tailored to costly experiments and safety-critical decisions. By distinguishing itself along these dimensions, justified abstention offers a unified normative layer that integrates technical tools into scientifically responsible practice.

Implications for Materials AI Practice

The adoption of justified abstention necessitates concrete changes across the materials AI community. For authors, three practices become essential. First, new models must implement at least one of the four abstention types within their inference pipelines, consistent with established advances in selective prediction and abstention-aware modeling [5, 8, 18]. Second, publications should report abstention rates alongside conventional metrics, transparently documenting the fraction of inputs declined under each criterion [8, 10]. Third, authors must justify the specific thresholds chosen for uncertainty, out-of-distribution scores, and risk functions, grounding them in domain-relevant utility considerations [7, 11].

Reviewers, in turn, acquire a new evaluative responsibility. They should routinely inquire whether a submitted model ever abstains and, if not, request explicit justification for the absence of refusal mechanisms. Manuscripts claiming universal applicability without addressing epistemic boundaries warrant heightened scrutiny, as unconditional prediction may signal an incomplete treatment of model limitations [2, 22].

At the community level, two initiatives are required. The development of abstention-aware benchmarks—datasets that explicitly reward principled refusal alongside accuracy—will accelerate progress [1, 8]. Additionally, the creation of standardized reporting templates for abstention behavior will foster comparability across studies. These changes collectively shift materials AI from a predict-at-all-costs culture to one that values epistemic humility, ultimately accelerating trustworthy discovery while safeguarding resources and safety.

Conclusion

This paper has articulated a theory of justified abstention for uncertain materials AI models, grounded in three core principles—competence boundary, risk threshold, and resource consideration—and operationalized through five explicit criteria and four distinct types. By defining justified abstention as the deliberate withholding of a prediction when its expected utility falls below that of abstaining, the framework addresses a conspicuous gap: the near-universal refusal of materials models to refuse. Abstention is repositioned not as model failure but as a scientifically responsible feature that enhances reliability, conserves experimental resources, and aligns AI behavior with the normative demands of high-stakes discovery. The consequences—reduced but higher-quality coverage, elevated trust, shifted decision responsibilities, and new learning opportunities—demonstrate that strategic silence can be more valuable than forced speech. Materials AI must therefore embrace justified abstention as an essential design element, moving beyond reflexive prediction toward epistemically mature systems capable of signaling their own limits. Only then can these models serve as true partners in the quest to understand and engineer the material world.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Jul 2024 Revised: 26 Aug 2024 Accepted: 27 Sep 2024
Published online: 18 January 2025

References

Omee SS, Fu N, Dong R, Hu M, Hu J. Structure-based out-of-distribution (OOD) materials property prediction: A benchmark study. npj Comput Mater. 2024;10(1):144.

Schuster D. Abstaining machine learning: Philosophical considerations. AI SOC. 2025;40(6):4213-33.

Geifman Y, El-Yaniv R. Selective classification for deep neural networks. Adv Neural Inf Process Syst. 2017;30:4878-87.

Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G, Mohd-Yusof J. Combating label noise in deep learning using abstention. arXiv preprint arXiv:1905.10964. 2019 May 27.

Geifman Y, El-Yaniv R. SelectiveNet: A deep neural network with an integrated reject option. In: International Conference on Machine Learning. New York, NY: PMLR; 2019:2151-9.

Lamy A, Zhong Z, Menon AK, Verma N. Noise-tolerant fair classification. Adv Neural Inf Process Syst. 2019;32:284-94.

Franc V, Prusa D, Voracek V. Optimal strategies for reject option classifiers. J Mach Learn Res. 2023;24(11):1-49.

Pugnana A, Perini L, Davis J, Ruggieri S. Deep neural network benchmarks for selective classification. arXiv preprint arXiv:2401.12708. 2024 Jan 23.

Gelbhart R, El-Yaniv R. The relationship between agnostic selective classification, active learning and the disagreement coefficient. J Mach Learn Res. 2019;20(33):1-38.

Varshney N, Mishra S, Baral C. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In: Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg, PA: Association for Computational Linguistics; 2022:1995-2002.

Narasimhan H, Menon AK, Jitkrittum W, Kumar S. Plugin estimators for selective classification with out-of-distribution detection. arXiv preprint arXiv:2301.12386. 2023 Jan 29.

Ly C, Nizinski C, McDonald IV LW, Chalifoux A, Hagen A. Out-of-distribution detection with non-parametric density estimation for models predicting processing history of uranium ore concentrates. Comput Mater Sci. 2025;259:114148.

Theunissen L, Mortier T, Saeys Y, Waegeman W. Evaluation of out-of-distribution detection methods for data shifts in single-cell transcriptomics. Brief Bioinform. 2025;26(3):bbaf239.

Cui P, Wang J. Out-of-distribution (OOD) detection based on deep learning: A review. Electronics. 2022;11(21):3500.

Jungo A, Doorenbos L, Da Col T, Beelen M, Zinkernagel M, Márquez-Neila P, et al. Unsupervised out-of-distribution detection for safer robotically guided retinal microsurgery. Int J Comput Assist Radiol Surg. 2023;18(6):1085-91.

Lv X, Li M, Chen J, Dong Z, Han S, Liao B. Out-of-distribution detection via LLM-guided outlier generation for text-attributed graph. In: Findings of the Association for Computational

Linguistics: ACL 2025. Stroudsburg, PA: Association for Computational Linguistics; 2025:19544-55.

Bitterwolf J, Meinke A, Augustin M, Hein M. Breaking down out-of-distribution detection: Many methods based on OOD training data estimate a combination of the same core quantities. In: International Conference on Machine Learning. New York, NY: PMLR; 2022:2041-74.

Wen B, Yao J, Feng S, Xu C, Tsvetkov Y, Howe B, et al. Know your limits: A survey of abstention in large language models. Trans Assoc Comput Linguist. 2025;13:529-56.

Gandouz M, Holzmann H, Heider D. Machine learning with asymmetric abstention for biomedical decision-making. BMC Med Inform Decis Mak. 2021;21(1):294.

Zhu Y, Nowak R. Efficient active learning with abstention. Adv Neural Inf Process Syst. 2022;35:35379-91.

Nguyen VL, Hullermeier E. Reliable multilabel classification: Prediction with partial abstention. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press; 2020;34(04):5264-71.

Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4.

Gardner M. Abstention at the border. Va Law Rev. 2019;105(1):63-126.

Schreuder N, Chzhen E. Classification with abstention but without disparities. In: Uncertainty in Artificial Intelligence. New York, NY: PMLR; 2021:1227-36.

Swaminathan A, Lopez I, Wang W, Srivastava U, Tran E, Bhargava-Shah A, et al. Selective prediction for extracting unstructured clinical data. J Am Med Inform Assoc. 2024;31(1):188-97.

Dvijotham K, Winkens J, Barsbey M, Ghaisas S, Stanforth R, Pawlowski N, et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. Nat Med. 2023;29(7):1814-20.

Feng J, Sondhi A, Perry J, Simon N. Selective prediction-set models with coverage rate guarantees. Biometrics. 2023;79(2):811-25.

Yoshikawa H, Okazaki N. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In: Findings of the Association for Computational Linguistics: EACL. Stroudsburg, PA: Association for Computational Linguistics; 2023:2017-28.

Dyer T, Lang M, Stice-Lawrence L. The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation. J Account Econ. 2017;64(2-3):221-45.