ANALYSIS

# In Risk Decisions, LLMs Should Emit Calibrated Probabilities, Not Answers: A Unified Framework for Calibration, Cost-Sensitive Thresholds and Deferral Capacity

HX Research

**Large language models are increasingly used for discrete risk decisions such as fraud screening, compliance review, clinical triage and ticket routing, and single-pass decision models that output option probabilities have recently appeared. Yet answers or verbalized confidence are often wired directly into rules engines, on the assumption that a more accurate model is safer, that calibration improves discrimination and that more referral to humans is more prudent. Here we argue that in risk decisions a language model should act as a provider of calibrated probabilities, not answers. Building on Bayesian decision and rejection theory, we unify cost-sensitive thresholds, deferral bands with reviewer error, calibration metrics, recalibration, conformal risk control and review capacity in six equations, and compare five classes of probability sources. Evidence and a reproducible simulation show that decision quality depends on whether probabilities can be reliably thresholded rather than on accuracy alone; that calibration leaves discrimination unchanged and pays off when costs change, when deferral is used and under drift; and that reviewer error and capacity shift the optimal deferral band. We give a calibrate–threshold–defer algorithm and five falsifiable judgements.**

2 November 2026 • HX Research

Nace.AI recently released Drex Max v1, a 27B dense "decision model", as open weights. It does not generate text; given a context and a set of named questions, it outputs the probability of each option in one forward pass, supporting multiple-choice, yes/no and ordinal-rating questions (according to the official model card and media reports)[1,2]. Such models have formed a recognizable category in recent months. A practitioner-oriented survey calls them "System 1 decision models"—models that map an input to a typed, structured decision in one forward pass—and lists calibration as a core theme[3]; a benchmark covering 105 tasks and 92,000 questions shows that larger System 1 models match or exceed fast closed LLMs in accuracy on non-agentic tasks, with better probability calibration and about a quarter of the latency (according to the paper)[4].

For risk decisions, the significance of this shift is not that another class of models exists, but that it surfaces a long-neglected issue: rules engines and human review need a probability that can be thresholded, not a passage of text. Bank fraud screening, pharmaceutical compliance pre-review, hospital triage and platform ticket routing are all trade-offs among the cost of false alarms, the cost of misses and the cost of review. A generated sentence such as "this transaction is risky" cannot be connected to a cost matrix; whether a verbalized "90% confidence" really means one error in ten also has to be tested. This need is long-standing in credit scoring: statistical classification methods in consumer credit scoring have been systematically reviewed[5], and a large benchmark compared dozens of classification algorithms and argued for evaluating scorecards with multiple performance measures rather than accuracy alone[6]; the methodological literature on medical prediction calls calibration "the Achilles heel of predictive analytics"[7].

Current practice has three gaps. First, probability sources are heterogeneous—verbalized confidence, log-probabilities of answer tokens, single-pass decision heads, attached calibration networks and sampling consistency—and differ greatly in calibration, granularity and cost, yet are often used as if they were the same thing. Second, calibration, thresholds and deferral are usually handled by different teams: data scientists report accuracy and ECE, business teams set thresholds by experience and operations teams set referral rates by headcount, with no shared metric. Third, counter-evidence is rarely

cited: one study finds that a decision model trails the best LLM on 14 of 15 social-science annotation tasks[8]; another finds that in prompt-injection detection a decision model's native probabilities do not consistently beat post-hoc calibrated classifier scores[9].

Our central claim is that in risk decisions an LLM should be connected to the system as a "probability provider", not an "answer provider"; decision quality is jointly determined by three things—whether probabilities can be reliably thresholded, whether the threshold comes from explicit costs, and whether deferral is constrained by human capacity. We make three contributions:

1. We place cost-sensitive thresholds, deferral bands with reviewer error, calibration metrics, recalibration, conformal risk control and review capacity in one notation through six numbered equations, and show their quantitative relationships.
2. We compare existing technical routes in three layers—probability source, post-processing and decision rule—setting supporting and opposing evidence side by side, and give each route's applicability boundaries and failure modes.
3. We give an executable calibrate–threshold–defer algorithm (Algorithm 1), a reproducible illustrative calculation, and five forward-looking judgements, each with a time horizon, evidence, a falsification condition and enterprise implications.

## Problem formalization and notation

### *Cost-sensitive binary decisions*

Let $y \in \{0, 1\}$ indicate whether a risk event occurs (1 is a bad case, such as fraud, a violation or a high-risk patient), x the observable information, $p(x) = \Pr(y = 1 \mid x)$ the true posterior and q(x) the probability given by the system. Let $C_{FN}$ be the cost of passing a bad case and $C_{FP}$ the cost of blocking a good case. For one case, the expected cost of "block" is $(1 - p)\cdot C_{FP}$ and of "pass" is $p\cdot C_{FN}$. The Bayes-optimal rule is

$$\textit{Block when } q > t^*, \textit{ with } t^* = C_{FP} / (C_{FP} + C_{FN}) \qquad (1)$$

Eq. (1) presupposes q = p, i.e. calibrated probabilities. With $C_{FN} = 10$ and $C_{FP} = 1$, $t^* = 1/11 \approx 0.091$, far below the default of 0.5 used by many systems. If q differs from p, each case misplaced on the wrong side of the threshold adds an excess cost of $|p\cdot C_{FN} - (1 - p)\cdot C_{FP}|$. This cost arises only near the threshold, so a global average calibration error can be small while decision loss is not. Zadrozny and Elkan studied score transformations precisely because classifier scores must first become accurate probabilities before they can be plugged into a cost matrix[10]; the notion of calibration itself goes back to Dawid's definition of the "well-calibrated Bayesian"[11].

### *Deferral bands with reviewer error*

Add a third action, "refer to human review", with fixed cost $C_D$ and reviewer error rate ε. The expected cost of referral is $C_D + \varepsilon\cdot[p\cdot C_{FN} + (1 - p)\cdot C_{FP}]$. Refer if and only if it is below the smaller of the two automatic actions:

$$\textit{Refer when } \min\{q\cdot C_{FN}, (1 - q)\cdot C_{FP}\} > C_D + \varepsilon\cdot[q\cdot C_{FN} + (1 - q)\cdot C_{FP}] \qquad (2)$$

With ε = 0, Eq. (2) reduces to the classical rejection rule, with deferral band $C_D/C_{FN} < q < 1 - C_D/C_{FP}$. With $C_D = 0.4$ the band is (0.04, 0.60). With ε > 0 the band narrows: reviewers also make mistakes, so referral itself carries a hidden cost. This is Chow's optimal error–reject trade-off[12]; selective classification for deep networks generalizes it to controlling risk at a given coverage[13].

### *Calibration metrics and recalibration*

Divide cases into B bins by q; let bin b hold a share $n_b/n$ of cases with mean probability $\bar{q}_b$ and observed bad rate $\bar{y}_b$. Expected calibration error and the Brier score are

$$ECE = \Sigma_b\,(n_b/n)\cdot|\bar{y}_b - \bar{q}_b|,\ Brier = (1/n)\cdot\Sigma_i\,(q_i - y_i)^2 \qquad (3)$$

The two common post-hoc recalibrations are monotone transforms of the model's raw logit z:

$$\textit{Temperature scaling: } q = \sigma(z / T);\ \textit{Platt scaling: } q = \sigma(a\cdot z + b) \qquad (4)$$

where σ is the logistic function. Temperature scaling has a single parameter and was first shown to be simple and effective on modern neural networks[14]; Platt scaling adds an intercept and can correct both overconfidence and an overall shift. Both are monotone and do not change the ordering of cases, and so do not change discrimination metrics such as AUC or KS. This point is central below. The Brier score was first used to verify probabilistic weather forecasts[15] and is a strictly proper scoring rule[16]; the binned form of ECE comes from Naeini and colleagues[17]. Binned estimates are themselves biased: Kumar and colleagues show that binning can underestimate calibration error[18], and Nixon and colleagues find ECE sensitive to the binning scheme[19]. Systematic comparison of isotonic regression and Platt scaling goes back to Niculescu-Mizil and Caruana[20], with generalizations such as Dirichlet calibration for the multi-class case[21]; Minderer and colleagues find that recent non-convolutional architectures are often better calibrated, and that model size and amount of pretraining do not fully explain calibration differences[22].

### *Conformal risk control*

Calibration gives a guarantee on average. If the business needs a finite-sample guarantee of the form "among automatically passed cases, the miss rate does not exceed α", conformal methods can be used. Let $\hat{R}_n(\lambda)$ be the empirical risk of threshold λ on the calibration set and B an upper bound on the loss of a single case. Conformal risk control selects[23]

$$\hat{\lambda} = inf\{ \lambda : [n \cdot \hat{R}_n(\lambda) + B] / (n + 1) \leqslant \alpha \} \quad (5)$$

If the calibration set and live cases are exchangeable, $\hat{\lambda}$ keeps expected live risk at or below α. This premise is exactly where it most easily fails in risk decisions, as discussed below under "Why calibration expires". Eq. (5) is the form of conformal risk control[24], built on the conformal prediction framework of Vovk and colleagues[25]; weighted extensions exist for covariate shift and non-exchangeable settings[26,27], but all require extra assumptions about the form of the shift.

### *Review capacity*

Let there be M reviewers, each with A minutes of effective review time per day, an average handling time of τ minutes per case and daily volume V. The upper limit on the referral rate is

$$\rho = Pr(refer) \leqslant \kappa = M \cdot A / (\tau \cdot V) \quad (6)$$

When the share covered by the deferral band of Eq. (2) exceeds κ, cases must be truncated in descending order of "expected saving from referral", not by simply narrowing the thresholds. Table 1 summarizes the notation.

**Table 1 | Main notation**

| Symbol | Meaning | Where introduced |
|---|---|---|
| y, x, p, q | Risk-event label, observable information, true posterior, system probability | Eq. (1) |
| $C_{FN}$, $C_{FP}$, t∗ | Cost of a miss, cost of a false alarm, Bayes threshold | Eq. (1) |
| $C_D$, ε | Fixed cost of referral, reviewer error rate | Eq. (2) |
| B, ECE, Brier | Number of bins, expected calibration error, Brier score | Eq. (3) |
| z, T, a, b | Raw logit, temperature, Platt parameters | Eq. (4) |
| α, $\hat{\lambda}$, $\hat{R}_n$ | Target risk, conformal threshold, empirical risk on the calibration set | Eq. (5) |
| M, A, τ, V, κ | Number of reviewers, daily effective time, handling time per case, daily volume, referral-rate cap | Eq. (6) |

## Analytical framework: probability source, post-processing and decision rule

We decompose "LLMs entering risk decisions" into three layers. The first is the probability source, answering "where does q come from"; the second is post-processing, answering "how is q corrected to approach p"; the third is the decision rule, answering "how does q become an action". Each layer has several technical routes with different failure modes.

There are five probability sources. Verbalized confidence asks the model to state a number after answering; it is cheapest but prompt-dependent. Answer-token probability reads the log-probabilities of option tokens at the answer position and normalizes over legal options; it needs white-box or logprob access. A single-pass decision head is a purpose-trained readout that outputs all option probabilities in one forward pass; Drex Max and Chinese Decision belong to this class[1,28]. An attached calibration network trains a small network on a frozen LLM, using internal features to predict whether an output is correct[29]. Sampling consistency samples the same question several times and uses the share of agreeing answers as confidence. Kadavath and colleagues found early that pretrained models are largely calibrated on the probability of their own answers in a suitable format[30]; the GPT-4 technical report shows that calibration worsens after post-training[31]; Tian and colleagues find that for models fine-tuned with human feedback, verbalized confidence is sometimes better calibrated than conditional probabilities[32], consistent with the counter-evidence below.

Post-processing has four classes: monotone recalibration such as temperature scaling, Platt scaling and isotonic regression; training-time calibration, such as calibration-aware reinforcement learning or entropy-targeted bilevel optimization[33,34]; conformal prediction and conformal risk control; and group-wise stratified calibration. Decision rules have three classes: a single threshold, a deferral band and a cascade, in which a small model or decision model handles cases first and uncertain ones go to a large model or to humans.

**Table 2 | Comparison of five probability sources (all evidence self-reported in papers or official materials, not independently reproduced)**

| Probability source | Mechanism | Main strength | Main problem | Suitable when | Representative evidence |
|---|---|---|---|---|---|
| Verbalized confidence | Model states a number after answering | Works on black boxes, cheap | Coarse, drifts with prompts, reflects "willingness to commit" more | API only, small samples | Ranking AUROC up to 0.956, but re-asking flips 4%–9% of decisions |
| Answer-token probability | Read option logits at the answer position | More tightly linked to correctness | Needs logprob access; post-training often makes it overconfident | Open models with logit access | Closer to an "evidence strength" signal than verbalized confidence |
| Single-pass decision head | Purpose-trained readout outputs all option probabilities | One pass, low latency, naturally thresholdable | Accuracy may trail the best LLM; needs in-domain recalibration | High volume, low latency, fixed options | ECE around 1% on some tasks; but trails the best LLM on 14 of 15 tasks |
| Attached calibration network | Predict correctness from frozen-model internal features | Base model unchanged | Needs labels; breaks when the base model changes | Stable base, readable internals | More reliable confidence on binary classification |
| Sampling consistency | Sample several times, count agreement | No internal access needed | Cost multiplies; amplifies shared misconceptions | Weak models, low-volume high-value decisions | Three-sample consistency AUROC about 0.77–0.79, below verbalized confidence |

The numbers in Table 2 come respectively from an analysis of verbalized confidence[35], comparisons of verbalized and internal confidence[36,37], evaluations of decision models[8,28] and an attached calibration network[29]. No source dominates on accuracy, calibration, granularity and cost at once. This is why the second and third layers are needed: the probability source only sets the starting point; usable decision probabilities are produced jointly by post-processing and decision rules.

## Mechanisms and route comparison

### *Why generative models' confidence cannot be wired directly into a threshold*

The first mechanism is that verbalized confidence reflects commitment, not correctness. Kumaran borrows a two-stage abstention paradigm from the neuroscience of perceptual decisions: the model answers and states its confidence, then decides whether to hand the answer to the user or abstain. According to the paper, verbalized confidence predicts "whether to commit" clearly better than "whether correct"; calibrated token log-probabilities show the opposite, coupling with discrimination of correctness; after removing the variance shared with log-probabilities, the residual of verbalized confidence still aligns with commitment while its link to correctness is near chance[37]. Zhang and colleagues compare verbalized and logit-based confidence across 30 models and 8 classification tasks and find that the two often diverge in association, magnitude and calibration; instruction-tuned models tend to report higher confidence, with larger confidence gaps and worse calibration; prompt design mainly changes the distribution of reported confidence[36].

The second mechanism is that post-training makes models more accurate and also more confident. Yaldiz and colleagues compare the calibration of supervised fine-tuning and reinforcement learning with verifiable rewards (RLVR) on decision tasks and find that RLVR improves task performance but produces severely overconfident models, while supervised fine-tuning is markedly better calibrated, including under distribution shift; their diagnosis is that the decision tokens in reasoning traces merely "extract" a decision already made and carry no confidence information (according to the paper)[33]. Bereket and Leskovec find that for binary predictions of stochastic outcomes GRPO induces overconfidence while PPO and RLOO are well calibrated, and that removing GRPO's group standard-deviation normalization fixes it[38]. Wang and colleagues further show that the confidence priors of off-the-shelf reasoning models concentrate on a few high values and persist through reinforcement learning, suppressing policy-gradient updates for rare confidence values[39].

The third mechanism concerns agents. Xuan and colleagues find that tool type produces a "confidence dichotomy": evidence tools such as web search systematically induce severe overconfidence because of noise in retrieved information, while verification tools such as code interpreters constrain reasoning through deterministic feedback and mitigate miscalibration (according to the paper)[40]. Evidence from medicine is more direct. One study adds linguistic uncertainty cues to MedMCQA, or removes the correct option so that the model ought to abstain; accuracy falls as uncertainty rises but confidence does not fall correspondingly, and "unsafe confident errors" increase markedly[41]. Another evaluation of 34 locally deployed clinical LLMs finds that providing clean evidence raises mean accuracy from 73.5% to 94.1% and cuts dangerous overconfidence from 8.0% to 1.6%, whereas standard RAG and agentic RAG do not reproduce this safety profile and extra reasoning compute brings limited gains (according to the

paper)[42]. The authors conclude that safety is not a passive consequence of scale.

Counter-evidence warns against the opposite extreme. Meyer and colleagues find that after auditing and correcting benchmark errors, verbalized confidence reported with the answer predicts correctness with AUROC of 0.956 and 0.937, clearly above three-sample consistency at 0.765 and 0.790; but when the same answer's confidence is re-elicited with equivalent prompts, scores change by 0.043 to 0.084 on average, flipping 4% to 9% of decisions at a threshold of 0.8 (according to the paper)[35]. Sun and colleagues call this a "score granularity gap": across 25 combinations of 9 LLMs and 3 benchmarks, single-shot verbalized confidence, correctly converted into class probabilities, ranks surprisingly well but takes only a handful of distinct values, and so offers only a few coarse thresholds[43]. For risk decisions the latter is decisive: a threshold such as $t* = 0.091$ cannot be implemented on a score that only reports 0.7, 0.8, 0.9 and 0.95. Early systematic evaluations also reported that verbalized confidence is generally too high[44].

### *Single-pass decision models: thresholdable, but not automatically more accurate*

Decision models solve the granularity problem directly: they output continuous probabilities over legal options. The Chinese Decision family compiles each question into a shared instruction and candidate prompt, reads the logits of candidate letters in one forward pass, normalizes over legal options and ships a temperature fitted on a validation set; according to the paper, its 4B model reaches 85.97% overall accuracy on the 307,900-decision CJ-Bench, 84.00% average accuracy across medicine, law and finance, and an expected calibration error of 1.00%[28]. In scam-call screening, JevLite fine-tunes Qwen3-4B with LoRA so that the temperature-scaled softmax over two answer-label logits is the probability of "scam"; according to the paper, on 577 turn-level decisions a three-seed ensemble reaches AUROC 0.974 and calibration error 0.052 with no false alarms on legitimate calls, at 64.5 milliseconds per decision—4.9 times faster than the same base fine-tuned to generate answers[45].

But the authors themselves state the boundaries. The JevLite paper notes that the gain comes from readout and calibration, not accuracy; a fine-tuned ModernBERT encoder is not significantly worse; recipe selection touched the test set; and all callers are synthetic[45]. Ibrahim and Zaki compare the first commercial decision model with 19 LLMs on 18 computational social science classification tasks; according to the paper, the decision model trails the best LLM for the task on 14 of 15 evaluated tasks, with a median gap of 11.6 macro-F1 points, but at about 44 times lower measured cost; its confidence is better calibrated than the verbalized confidence of 16 LLMs, but three frontier models have lower median calibration error (0.066 versus 0.157); on an empathy task in peer-support conversations it reports high confidence while performing near chance[8]. Their recommendation is to put the decision model first in the pipeline: routing low-confidence cases to an LLM matches or exceeds LLM-only performance at a quarter to a half of the cost.

Evidence from security tasks is similar. Lan and colleagues evaluate prompt-injection detection on 47,611 held-out samples; according to the paper, an LLM judge has the highest discrimination (AUROC 0.897) and the decision model is second (0.871); the decision model's native probabilities do not consistently beat post-hoc calibrated classifier scores; temperature scaling improves calibration but not discrimination; and at a 1% miss-rate target every "gate + review" system needs an escalation rate above 91.8%[9]. Risk teams should heed the last point: when the miss target is very strict and discrimination is insufficient, no calibration can avoid sending almost every case to humans. System1-Arena also shows that in closed-loop environments large open LLMs have the highest success rate (up to 32.9%) while small System 1 models rarely succeed (according to the paper)[4].

Taken together, decision models change "whether probabilities can be implemented", not "the ceiling of discrimination". The fact that the monotone recalibration of Eq. (4) does not change AUC agrees with the empirical finding that temperature scaling "improves calibration but not discrimination"[9].

### *Why calibration "expires": transfer, configuration drift and compression*

Calibration parameters are fitted under a particular distribution and inference configuration. Jin and colleagues point out that post-hoc temperature scaling is inherently domain-dependent, and a temperature fitted in one domain does not generalize to others[34]. Afrin and Shihab test, across 13 LLM judges, 8 domains and 1,176 pre-registered transfers, whether the accuracy gap between source and target domains predicts recalibration failure; according to the paper, the same gap can lead to opposite transfer outcomes, and the leakage-free correlation is only 0.25, with a 95% confidence interval that includes 0[46]. Ovadia and colleagues' large-scale evaluation showed early that under dataset shift the calibration of all classes of uncertainty methods degrades as shift increases[47].

For LLMs there is an often-overlooked source of shift beyond the data distribution: the inference

configuration. Zhu and colleagues call changes of prompt template, decoding temperature and weight quantization "configuration shift"; across 9 models, 4 datasets and 4 nonconformity scores they find that configuration shift persistently erodes the validity of conformal prediction, with empirical coverage often below target while prediction-set size barely changes (according to the paper)[48]. Tong and colleagues find across 12 models that compression often decouples accuracy from uncertainty, that larger models absorb compression-induced uncertainty better, and that uncertainty inflation is often threshold-like rather than gradual[49]. This means that a release that "only switched the quantized version" or "only changed the prompt template" can invalidate the guarantee of Eq. (5) even if accuracy regression tests pass.

Group-level problems are subtler. Salem and colleagues find that under mild shifts in group composition standard conformal risk control violates the risk budget in up to 47% of trials; their hierarchical group-conditional method achieves zero violations on some models, but at a participation cost of 22 to 37 percentage points relative to the global method—that is, more abstention (according to the paper)[50]. SHIFT-Jev offers a monitoring, abstention and continual-adaptation pipeline for decision models under distribution shift; according to the paper, KL-anchored fine-tuning plus temperature scaling cuts calibration error under shift from 0.289 to 0.087; at a 10% target risk, a learned uncertainty score with a certified threshold exceeds the target in 3 to 5 of 27 tests, versus 5 to 15 for the equally certified maximum probability score and 9 to 14 for conformal prediction sets; and a chi-square test on predicted labels detects 95% to 99% of class-prior shifts[51]. The last point matters especially for risk businesses: a change in the bad rate is the shift that confidence-based monitoring most easily misses. Detection and correction of class-prior shift have mature methods: Saerens and colleagues adjust classifier outputs to new priors with EM[52], and Lipton and colleagues estimate and correct label shift using the confusion matrix of a black-box predictor[53].

### *Deferral and humans: capacity, fatigue and bias*

Eqs. (2) and (6) put people into the decision. The corresponding direction in machine learning is "learning to defer". Alves and colleagues propose DeCCaF for cost-sensitive fraud detection, using supervised learning to estimate each human expert's error probability and then constraint programming to globally minimize error cost under workload constraints[54], and release the FiFAR dataset with predictions from 50 synthetic fraud analysts and a realistic definition of human capacity[55]. Zhang and colleagues drop the assumption of constant human performance, model human performance declining with cumulative workload using psychological fatigue curves, and optimize the collaboration budget in a constrained Markov decision process[56]. Early formalizations of learning to defer come from Madras and colleagues[57], and Mozannar and Sontag give a consistent surrogate loss[58].

People not only tire; they are also changed by the cases deferred to them. Pesenti and colleagues find that standard learning-to-defer policies disproportionately defer the minority class under class imbalance; a user study with 226 participants shows that, faced with highly imbalanced deferred sets, participants classify the majority class less accurately[59]. Research on agent oversight points the same way: reviewers agree only moderately on what counts as a risky action (Fleiss $\kappa = 0.52$), and once fatigue is considered safety is an inverted U in the escalation rate[60]. Together these findings show that a deferral band cannot be drawn from model uncertainty alone; it must also account for the composition of cases that reach human hands. Human–AI collaboration research also finds that AI explanations increase acceptance of AI recommendations regardless of whether they are correct, so human–AI teams do not always outperform either alone[61].

On "use uncertainty or learn a deferral model", Wundram and Baumgartner find on ophthalmology data that uncertainty quantification can be more robust than a supervised deferral model under out-of-distribution inputs[62]. On deployment, a customer-service system on an enterprise business-process platform trains a critic on agents' low-cost "accept or edit the suggestion" feedback to calibrate abstention, auto-executing only high-confidence steps and handing uncertain ones to the agent; according to the paper it automated 45% of conversations and cut average handling time by 39% without lowering service quality[63]. A traffic-crash severity grading study shows the typical trade-off of selective deferral: retaining 70% of cases for automatic classification raises macro-F1 on the automatic part from 0.522 to 0.573, while the deferred part is treated as candidates for human review and is not evaluated further[64]. That last clause deserves attention: many papers report metrics for the "automatic part", while true outcomes of the deferred part go unmeasured. The maximum softmax probability is the classic baseline for detecting misclassified and out-of-distribution examples[65].

### *Division of labour with the rules engine: probabilities propose, rules decide*

In risk decision systems the LLM is not the only decision-maker. Akhtyamov discusses self-evolving agents in credit workflows and argues that only by confining self-modification to the runtime execution layer (instruction text, tool-calling logic and primitive composition) while keeping model weights fixed can each adaptation become a diff with a stated reason and tests; according to the paper, in simulation a loop with an admission gate admitted 144 of 7,449 candidate changes, none of which worsened error on held-out history[66]. From a reproducibility angle, Zhou and colleagues' survey notes that LLM agent workflows in financial AI show batch-dependent output divergence and trajectory drift and need a layered evaluation framework[67].

This evidence supports a clear division of labour: probabilities are produced by the model and then calibrated and certified; thresholds are derived from the cost matrix and kept under version control as configuration; final actions are executed by the rules engine, whose rules can override probabilities—for example regulatory hard requirements, blacklists and limit caps; referred cases carry the probability, the rules triggered and the evidence, and review outcomes flow back into the calibration set. Early research on LLMs for credit scoring also suggests generalization potential on some credit tasks alongside bias risks[68]. Regulatory frameworks set similar expectations: the revised model risk management guidance issued by US federal banking agencies in April 2026 (SR 26-2, superseding SR 11-7) takes a risk-based approach, tailoring the management of model development, validation and use to the model risk profile (according to the official document)[69].

**Table 3 | Comparison of post-processing and decision-rule routes**

| Route | Mechanism | Main benefit | Main risk | Suitable when | Not suitable when |
|---|---|---|---|---|---|
| Temperature scaling | One-parameter monotone transform | Simple, little data needed | Domain-dependent; does not fix overall shift | Same domain, limited samples | Bad rate changes substantially |
| Platt / isotonic | Two-parameter or nonparametric monotone transform | Fixes shift and overconfidence together | Overfitting on small samples (isotonic) | Enough in-domain labels | Very few labels |
| Training-time calibration | Calibration-aware RL, entropy targets | More stable across domains | Requires retraining; costly | In-house or fine-tunable models | Closed API only |
| Conformal risk control | Finite-sample risk guarantee | Auditable upper bound | Relies on exchangeability; group violations | Calibration set matches live distribution | Frequent configuration or population changes |
| Deferral band | Refer per Eq. (2) | Lowers expected cost | Human capacity, fatigue and bias | Professional review team available | High reviewer error rate |
| Cascade | Decision model screens first; LLM or human decides later | Large cost reduction | Cascade thresholds need in-domain certification | High-volume decisions, long-tail hard cases | Decision model badly miscalibrated in domain |

## Method: the calibrate–threshold–defer procedure

The analysis above combines into an executable procedure. Given a probability source and a labelled in-domain sample, Algorithm 1 outputs an implementable, auditable decision configuration: a recalibration map, a cost threshold, a deferral band and a certified risk bound.

**Algorithm 1** Calibrate–Threshold–Defer (CTD)

**Input:** Labelled in-domain sample D (split by time into fitting set $D_1$, calibration set $D_2$ and out-of-time test set $D_3$); raw scores z from the probability source; costs $C_{FN}$, $C_{FP}$, $C_D$; an estimate of reviewer error ε; capacity parameters M, A, τ, V; target risk α; group partition G (such as product line, region, channel)

**Output:** Recalibration map g; threshold $t^*$; deferral band; referral rate $\rho \leqslant \kappa$; conformal threshold $\hat{\lambda}$ and the configuration fingerprint it applies to; expected cost, ECE and per-group risk on the out-of-time test set

1: Freeze the inference configuration (model version, prompt template, decoding parameters, quantization) and compute its fingerprint; all subsequent conclusions hold only for that fingerprint

2: On $D_1$, count the distinct score values; if fewer than 20, declare "insufficient granularity" and switch probability source or aggregate multiple queries

3: Fit g on $D_1$ (temperature scaling, Platt or isotonic regression, chosen by negative log-likelihood on $D_2$) to obtain $q = g(z)$

4: Compute $t^*$ by Eq. (1), the deferral band by Eq. (2) and κ by Eq. (6); if the band covers more than κ, truncate in descending order of expected saving $= \min\{q \cdot C_{FN}, (1-q) \cdot C_{FP}\} - C_D - \varepsilon \cdot [q \cdot C_{FN} + (1-q) \cdot C_{FP}]$

5: Compute $\hat{\lambda}$ on $D_2$ by Eq. (5) and re-check each group $g \in G$ separately, applying a Bonferroni correction when there are many groups

6: On $D_3$, report expected cost, AUC, ECE, Brier, referral rate and per-group risk; compare with two baselines, "raw score + 0.5 threshold" and "raw score + empirically optimal threshold"

7: After launch, use three triggers—a change of configuration fingerprint, a chi-square test on the distribution of predicted labels, and a rolling estimate of calibration error; any trigger returns to step 3

**Complexity and conditions:** Fitting and threshold computation are O(n log n); requires in-domain labels with acceptable label delay; requires approximate exchangeability between the calibration set and live traffic; when bad cases are very rare (for example below 0.5%), estimates of ECE and group risk have high variance, and Brier and cost should be reported first

Three design choices deserve explanation. Step 1 treats the inference configuration as part of calibration, following directly from research on configuration shift and compression[48,49]: the same model with a different prompt template or quantized version is a new system that needs recalibration. The granularity check in step 2 comes from the score granularity gap[43]. Step 7 lists a chi-square test on predicted labels as a separate trigger because class-prior shift is the change that confidence-based monitoring most easily misses[51].

## Illustrative calculation

We use one set of illustrative data to show how Eqs. (1)–(6) interact. The data are entirely simulated and do not correspond to any real business; the scripts src/sim.py and src/figs.py use fixed random seeds and are reproducible.

Setup: the bad rate is 5%; the true posterior comes from a two-Gaussian model with separation 2.2; the model's raw logit is twice the true posterior logit plus 1, so it is both overconfident and biased upwards. The fitting set has 5,000 cases and the test set 20,000. Costs are $C_{FN}$ = 10, $C_{FP}$ = 1 and $C_D$ = 0.4, reviewer error $\varepsilon$ = 5% and the capacity cap $\kappa$ = 10%. By Eq. (1), $t^* = 1/11 \approx 0.091$; by Eq. (2) with $\varepsilon = 0$, the deferral band is (0.04, 0.60).

**Table 4 | Illustrative calculation: expected cost per 1,000 cases on the test set under five configurations (illustrative data)**

| Configuration | ECE | Expected cost per 1,000 | Referral rate |
|---|---|---|---|
| Raw probability, threshold 0.5 | 0.020 | 256.8 | 0 |
| Raw probability, cost threshold $t^*$ | 0.020 | 198.8 | 0 |
| Temperature-scaled, cost threshold $t^*$ | 0.011 | 191.8 | 0 |
| Platt-calibrated, cost threshold $t^*$ | 0.002 | 184.2 | 0 |
| Platt-calibrated + deferral band (capacity 10%) | 0.002 | 162.5 | 10.0% |

Table 4 yields four conclusions. First, the choice of threshold matters more than the model: with the same raw probabilities, moving from the default 0.5 to the cost-derived $t^*$ lowers expected cost per 1,000 cases from 256.8 to 198.8, about 23%. Second, a small calibration error does not mean small decision loss: the raw probabilities have an ECE of only 0.020 and look "already accurate", yet applying $t^*$ to them costs about 8% more than after calibration. With a 5% bad rate most cases fall in bins near 0 whose errors are tiny, pulling ECE down, while decision loss concentrates near $t^*$ (Fig. 1). Third, temperature scaling fixes only overconfidence, not the overall shift, and so is inferior to Platt scaling; the calibrated cost of 184.2 is essentially equal to 184.6 under the true posterior, indicating that calibration is near its limit. Fourth, referring within 10% capacity lowers cost by about another 12%, to 162.5.

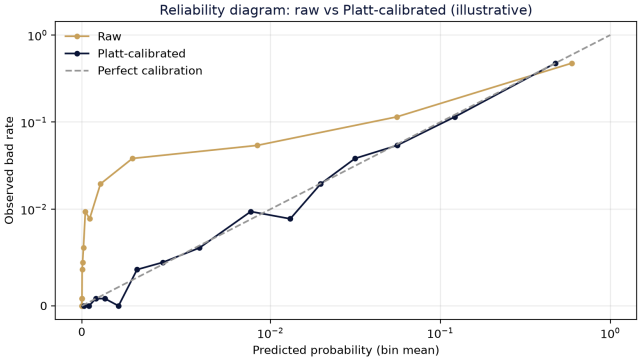



**Fig. 1 | Reliability diagram.** Both axes are on log scales. Raw probabilities are clearly above the observed bad rate in the high-score range; after Platt calibration they lie close to the diagonal. Illustrative data computed by the Great Unicorn Intelligence Research Team.

One counter-intuitive contrast deserves emphasis: if instead of calibrating we search the raw probabilities directly for the empirically optimal threshold, it is about 0.025, with cost about 184.9—almost the same as after calibration (Fig. 2). This agrees with the monotonicity in Eq. (4): with fixed costs, a stable distribution and a single binary decision, calibration is equivalent to "choosing

another threshold". The independent value of calibration appears in three situations: when the cost matrix changes and a new threshold must be computed directly from Eq. (1); when the deferral band of Eq. (2) is needed, since its two ends depend on the absolute value of the probability; and when probabilities must be combined with others (rule hit rates, other models) or compared across groups.

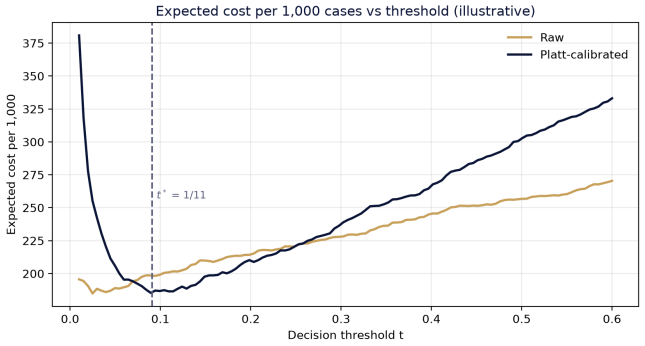



**Fig. 2 | Expected cost per 1,000 cases as a function of the decision threshold.** The cost minimum of the raw probabilities lies away from t∗; after calibration the minimum falls near t∗. Illustrative data computed by the Great Unicorn Intelligence Research Team.

Finally, capacity. Fig. 3 relaxes the capacity cap from 0 to 25%. Cost falls as capacity grows, reaches a minimum of about 155.1 at about 16%, then rises slightly and settles at about 157.4. The reason is that the deferral band drawn by Eq. (2) with ε = 0 covers about 18% of cases, but reviewers have a 5% error rate, so for cases near the edge of the band the true cost of referral already exceeds that of an automatic decision. Including ε in Eq. (2) gives the narrower band directly. This is a counter-example to the intuition that "more referral is safer".

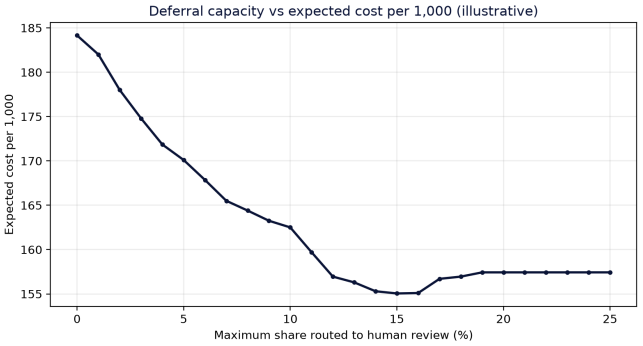



**Fig. 3 | Deferral capacity and expected cost per 1,000 cases.** Cost is lowest at about 16%, then rises slightly because reviewers make errors. Illustrative data computed by the Great Unicorn Intelligence Research Team.

## Forward-looking judgements and research agenda

The following judgements are the opinions of the Great Unicorn Intelligence Research Team, not statements of fact. Each gives a time horizon, evidence, a falsification condition and enterprise implications.

**Table 5 | Five forward-looking judgements (opinion)**

| No. | Judgement | Horizon | Main evidence | Falsified if | Enterprise implication |
|---|---|---|---|---|---|
| 1 | In high-volume risk decisions with fixed options, single-pass decision models or answer-token readout will replace "generated answer + verbalized confidence" as the default interface | 12–18 months | Calibration and latency advantages of decision models[4,45]; granularity gap of verbalized confidence[43] | Mainstream generative APIs widely offer continuous, thresholdable confidence outputs with third-party-verified calibration | Make "thresholdable continuous probability" a hard selection requirement |
| 2 | Cascades of decision models and frontier LLMs, rather than either alone, will become the mainstream deployment form | 6–12 months | Routing low-confidence cases to an LLM matches performance at a quarter to a half of the cost[8]; conformal cascades[70] | A single model dominates on both accuracy and cost, and cascades are widely abandoned in production | Certify cascade thresholds in domain and log the referral rate at each stage |
| 3 | Inference configuration fingerprints (model, prompt, decoding, quantization) will be bound to calibration parameters and enter model-risk change management | 12–24 months | Configuration shift erodes conformal guarantees[48]; compression decouples uncertainty[49] | Large-scale studies show common configuration changes generally have negligible effect on calibration and coverage | Any configuration change triggers recalibration and out-of-time backtesting |
| 4 | Referral policies will shift from "proportional to uncertainty" to joint optimization over capacity, reviewer error and case mix | 12–24 months | Fatigue-aware deferral[56]; deferral bias[59]; deferral under cost and capacity constraints[54] | Production evaluations show the gap between simple uncertainty thresholds and joint optimization is generally below 2% | Estimate each reviewer's error rate and write ε into Eq. (2) |
| 5 | Calibration constraints will become standard in post-training: reinforcement learning for decision tasks will include a calibration loss or calibration-aware reward by default | 12–24 months | RLVR and GRPO induce overconfidence[33,38]; concentrated confidence priors[39] | New-generation reasoning models without special treatment are generally better calibrated on decision tasks than supervised fine-tuned models | Require vendors to provide per-domain calibration reports |

Around these judgements we see four research questions worth investment: calibration evaluation at low bad rates, where ECE is systematically low and cost or local calibration near the threshold should

take priority; online recalibration under label delay, since performance windows in risk businesses are often measured in months; tracking the true outcomes of deferred cases, rather than reporting only metrics for the automatic part; and using rules-engine hit information as a covariate for calibration, studying the joint calibration of probabilities and rules.

## Where common industry views may be wrong

First, "a more accurate model means we can use its answers directly". Decisions need thresholdable probabilities; even for the same probabilities, the difference between a default threshold of 0.5 and the cost threshold produced about a 23% cost difference in the illustrative calculation.

Second, "a low ECE means calibration is done". At low bad rates ECE is pulled down by the many cases near 0 while decision loss concentrates near the threshold; local calibration near the threshold and expected cost should be reported as well.

Third, "calibration improves discrimination". Monotone recalibration does not change ordering, and temperature scaling improving calibration but not discrimination has been shown empirically[9]. The value of calibration appears when costs change, under deferral and in combination.

Fourth, "just ask the model how sure it is". Verbalized confidence reflects whether the model is "willing to commit" more than whether it is "right"[37], and it is coarse and prompt-sensitive[35,43].

Fifth, "more referral to humans is safer". Reviewers make errors and tire, and the composition of the deferred set affects their judgement[56,59]; in the illustrative calculation, cost rises beyond a certain share.

Sixth, "calibrate once, valid for good". Changes in model version, prompt template, quantization and population all make calibration expire[48,50].

## Limitations

This analysis has four limitations. First, most cited studies are preprints from the past 12 months whose numbers are self-reported and not yet peer-reviewed or independently reproduced; some decision-model results come from developers or use synthetic data. Second, the illustrative calculation uses an idealized two-Gaussian data-generating process, and costs, reviewer error and capacity are assumed values used only to show mechanisms; real risk businesses have label delay, reject inference and feedback loops that make calibration harder. Third, we mainly discuss binary decisions; threshold and deferral design for multi-class and ordinal decisions is more complex. Fourth, newly released models such as Drex Max have no independent in-domain evaluation yet; we cite only their official descriptions and make no judgement about their performance in any risk business.

## Conclusion

To return to the central claim: in risk decisions, LLMs should be connected to the system as probability providers. Decision quality is jointly determined by whether probabilities can be reliably thresholded, whether the threshold comes from explicit costs and whether deferral is constrained by human capacity. The verbalized confidence of generative models ranks reasonably well but reflects commitment more than correctness and is coarse and prone to drift; single-pass decision models solve granularity and latency but are not automatically more accurate, need in-domain recalibration and are often paired with LLMs in cascades; calibration does not change discrimination, and its value appears when costs change, under deferral and under drift; referral must account for capacity, reviewer error and case mix together.

In Great Unicorn Intelligence's practice, this division of labour is reflected in the design of two products. According to the company website, the SAM data analysis and scorecard modelling platform uses a supervisor agent to orchestrate seven specialist agents through a fourteen-step modelling pipeline with fourteen named human gates, and a change at any step invalidates and reruns downstream steps[71]; the governance layer of the Enterprise Agent Platform Harness includes a rules engine (policy as code, deciding whether a business action is allowed), permissions and authorization re-checked on every call and a replayable audit trail, holding that "the model only proposes, the system decides"[72]. This matches the framework here: probabilities are provided by the model and calibrated, thresholds and rules are controlled configuration, and final authority stays with rules and named people.

Key takeaways:

1. Have the model output probabilities, not answers, and first check whether the score's granularity can support the thresholds you need.
2. Derive the threshold from the cost matrix: $t* = C_{FP} / (C_{FP} + C_{FN})$; do not default to 0.5.
3. At low bad rates do not look only at ECE; look at calibration near the threshold and at expected cost.
4. Write reviewer error and capacity into the deferral band; more referral is not always better.
5. Bind the inference configuration fingerprint to calibration parameters; any configuration change triggers recalibration.

## References


1. Nace.AI. Drex Max v1 (model card). Hugging Face, https://huggingface.co/nace-ai/drex-max-v1 (2026).
2. Razzaq, A. Nace AI Open-Sources Drex Max v1: A 27B Decision Model That Runs Locally on Apple Silicon. MarkTechPost, https://www.marktechpost.com/2026/11/01/nace-ai-open-sources-drex-max-v1-local-inference/ (2026).
3. Roitman, H. The Hitchhiker's Guide to System 1 Models: Fast Decision-Making Beyond LLMS. Preprint at https://arxiv.org/abs/2610.33775 (2026).
4. Chen, T., Liu, X., An, R., Wang, Z. & Zhou, M. System1-Arena: Which Decisions Can Be Left to System 1?. Preprint at https://arxiv.org/abs/2610.21244 (2026).
5. Hand, D. J. & Henley, W. E. Statistical Classification Methods in Consumer Credit Scoring: A Review. *J. R. Stat. Soc. A* **160**, 523–541 (1997); https://doi.org/10.1111/j.1467-985X.1997.00078.x.
6. Lessmann, S., Baesens, B., Seow, H. V. & Thomas, L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* **247**, 124–136 (2015); https://doi.org/10.1016/j.ejor.2015.05.030.
7. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019); https://doi.org/10.1186/s12916-019-1466-7.
8. Ibrahim, H. & Zaki, Y. Evaluating Decision Models for Text Annotation in Computational Social Science. Preprint at https://arxiv.org/abs/2609.24574 (2026).
9. Lan, Q., Kaul, A., Pandurangan, V. & Sanyal, I. From Detection to Decision: Calibration and Deferral for Prompt-Injection Detection. Preprint at https://arxiv.org/abs/2610.20907 (2026).
10. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 694–699 (ACM, 2002); https://doi.org/10.1145/775047.775151.
11. Dawid, A. P. The Well-Calibrated Bayesian. *J. Am. Stat. Assoc.* **77**, 605–610 (1982); https://doi.org/10.1080/01621459.1982.10477856.
12. Chow, C. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory* **16**, 41–46 (1970); https://doi.org/10.1109/TIT.1970.1054406.
13. Geifman, Y. & El-Yaniv, R. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems* **30** (2017); https://arxiv.org/abs/1705.08500.
14. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On Calibration of Modern Neural Networks. In *Proc. 34th International Conference on Machine Learning* 1321–1330 (PMLR, 2017); https://arxiv.org/abs/1706.04599.
15. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950); https://doi.org/10.1175/1520-0493(1950)078<;0001:VOFEIT>2.0.CO;2.
16. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007); https://doi.org/10.1198/016214506000001437.
17. Pakdaman Naeini, M., Cooper, G. & Hauskrecht, M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proc. AAAI Conference on Artificial Intelligence* **29**, 2901–2907 (2015); https://doi.org/10.1609/aaai.v29i1.9602.
18. Kumar, A., Liang, P. & Ma, T. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems* **32** (2019); https://arxiv.org/abs/1909.10155.
19. Nixon, J. et al. Measuring Calibration in Deep Learning. In *Proc. CVPR Workshops* (2019); https://arxiv.org/abs/1904.01685.
20. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proc. 22nd International Conference on Machine Learning* 625–632 (ACM, 2005); https://doi.org/10.1145/1102351.1102430.
21. Kull, M. et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems* **32** (2019); https://arxiv.org/abs/1910.12656.
22. Minderer, M. et al. Revisiting the Calibration of Modern Neural Networks. In *Advances in Neural Information Processing Systems* **34** (2021); https://arxiv.org/abs/2106.07998.
23. Angelopoulos, A. N. & Bates, S. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. Preprint at https://arxiv.org/abs/2107.07511 (2021).
24. Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L. & Schuster, T. Conformal Risk Control. In *Proc. International Conference on Learning Representations* (2024); https://arxiv.org/abs/2208.02814.
25. Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World* (Springer, 2005); https://doi.org/10.1007/b106715.
26. Tibshirani, R. J., Barber, R. F., Candès, E. J. & Ramdas, A. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems* **32** (2019); https://arxiv.org/abs/1904.06019.
27. Barber, R. F., Candès, E. J., Ramdas, A. & Tibshirani, R. J. Conformal prediction beyond exchangeability. *Ann. Stat.* **51**, 816–845 (2023); https://doi.org/10.1214/23-AOS2276.
28. Yan, Q. Chinese Decision: A Family of Language Models for Single-Pass Probabilistic Decision-Making. Preprint at https://arxiv.org/abs/2610.23170 (2026).
29. Wang, J. et al. ProbPlug: A Plugin Uncertainty Network for Reliable Confidence in LLM Binary Classification. Preprint at https://arxiv.org/abs/2609.10122 (2026).
30. Kadavath, S. et al. Language Models (Mostly) Know What They Know. Preprint at https://arxiv.org/abs/2207.05221 (2022).
31. OpenAI. GPT-4 Technical Report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
32. Tian, K. et al. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proc. Conference on Empirical Methods in Natural Language Processing* (2023); https://arxiv.org/abs/2305.14975.
33. Yaldiz, D. N. et al. Balancing Classification and Calibration Performance in Decision-Making LLMs via Calibration Aware Reinforcement Learning. Preprint at https://arxiv.org/abs/2601.13284 (2026).
34. Jin, R., Wang, Z., Dai, Z., Xiao, J. & Hou, B. Beyond Post-Hoc Temperature Scaling: Bilevel Optimization for LLM Calibration. Preprint at https://arxiv.org/abs/2608.07419 (2026).
35. Meyer, L., Rossi, S., Chen, W., Laurent, T. & Li, Y. An Analysis of Training-Free Self-Reported Confidence in Language Models. Preprint at https://arxiv.org/abs/2609.20541 (2026).
36. Zhang, H. et al. When Linguistic and Internal Confidence Diverge in Large Language Models. Preprint at https://arxiv.org/abs/2608.28382 (2026).
37. Kumaran, D. Reported Confidence in LLMs Tracks Commitment More Than Correctness. Preprint at https://arxiv.org/abs/2606.29490 (2026).
38. Bereket, M. & Leskovec, J. Uncalibrated Reasoning: GRPO Induces Overconfidence for Stochastic Outcomes. Preprint at https://arxiv.org/abs/2508.11800 (2025).
39. Wang, S. et al. On the Pitfalls of Verbalized Confidence Priors for Calibrating Large Reasoning Models. Preprint at https://arxiv.org/abs/2609.32470 (2026).
40. Xuan, W. et al. The Confidence Dichotomy: Analyzing and Mitigating Miscalibration in Tool-Use Agents. Preprint at https://arxiv.org/abs/2601.07264 (2026).
41. Tahermazandarani, M., Mahmood, A., Islam, F. & Sheng, Q. Z. When Confidence Fails: Overconfidence in LLMs under Uncertainty and Missing Clinical Information. Preprint at https://arxiv.org/abs/2608.09080 (2026).
42. Wind, S. et al. Safety and accuracy follow different scaling laws in clinical large language models. Preprint at https://arxiv.org/abs/2605.04039 (2026).
43. Sun, A., Sun, T. & Geng, J. The Score Granularity Gap in Black-Box LLM Classification: A Comparative Study of Confidence Constructions. Preprint at https://arxiv.org/abs/2606.22179 (2026).
44. Xiong, M. et al. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. Preprint at https://arxiv.org/abs/2306.13063 (2023).
45. Ren, S. et al. Open-Jev Judgments on CallScreenBench: Calibrated One-Pass Scam Screening with a Small Language Model. Preprint at

https://arxiv.org/abs/2609.23959 (2026).
46. Afrin, F. & Shihab, I. F. When Accuracy Gaps Fail to Certify: Auditing Cross-Domain Recalibration of LLM Judges. Preprint at https://arxiv.org/abs/2609.27954 (2026).
47. Ovadia, Y. et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Advances in Neural Information Processing Systems* **32** (2019); https://arxiv.org/abs/1906.02530.
48. Zhu, Y. et al. Conformalized Large Language Models under Configuration Shift. Preprint at https://arxiv.org/abs/2608.01460 (2026).
49. Tong, Y. et al. Does Compression Preserve Uncertainty? A Unified Benchmark for Quantized and Sparse LLMs via Conformal Prediction. Preprint at https://arxiv.org/abs/2606.01850 (2026).
50. Salem, M., Böhm, L., Pontes, D. & Ferrugem, A. Hierarchical Group-Conditional Conformal Risk Control for Selective Prediction in Language Models. Preprint at https://arxiv.org/abs/2607.24562 (2026).
51. Saeed, M. A. et al. SHIFT-JEV: Certified Abstention and Gain-Gated Continual Adaptation for LLM-Based Decision Models under Distribution Shift. Preprint at https://arxiv.org/abs/2610.34329 (2026).
52. Saerens, M., Latinne, P. & Decaestecker, C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Comput.* **14**, 21–41 (2002); https://doi.org/10.1162/089976602753284446.
53. Lipton, Z. C., Wang, Y. & Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proc. 35th International Conference on Machine Learning* (PMLR, 2018); https://arxiv.org/abs/1802.03916.
54. Alves, J. V. et al. Cost-Sensitive Learning to Defer to Multiple Experts with Workload Constraints. Preprint at https://arxiv.org/abs/2403.06906 (2024).
55. Alves, J. V. et al. FiFAR: A Fraud Detection Dataset for Learning to Defer. Preprint at https://arxiv.org/abs/2312.13218 (2023).
56. Zhang, Z., Nguyen, C. C., Rosewarne, D., Wells, K. & Carneiro, G. Fatigue-Aware Learning to Defer via Constrained Optimisation. Preprint at https://arxiv.org/abs/2604.00904 (2026).
57. Madras, D., Pitassi, T. & Zemel, R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems* **31** (2018); https://arxiv.org/abs/1711.06664.
58. Mozannar, H. & Sontag, D. Consistent Estimators for Learning to Defer to an Expert. In *Proc. 37th International Conference on Machine Learning* (PMLR, 2020); https://arxiv.org/abs/2006.01862.
59. Pesenti, D., Bogani, A., Teso, S. & Pugnana, A. Too Much of the Same: From Algorithmic to Human Bias in Learning to Defer. Preprint at https://arxiv.org/abs/2608.28050 (2026).
60. Turan, E. Oversight Has a Capacity: Calibrating Agent Guards to a Subjective, Fatiguing Human. Preprint at https://arxiv.org/abs/2606.08919 (2026).
61. Bansal, G. et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* 1–16 (ACM, 2021); https://doi.org/10.1145/3411764.3445717.
62. Wundram, A. M. & Baumgartner, C. F. Is Uncertainty Quantification a Viable Alternative to Learned Deferral?. Preprint at https://arxiv.org/abs/2508.02319 (2025).
63. Borovkov, N. et al. Learning Selective LLM Autonomy from Copilot Feedback in Enterprise Customer Support Workflows. Preprint at https://arxiv.org/abs/2604.23855 (2026).
64. Chhetri, G., Baitullah, A. & Das, S. Toward Auditable and Calibrated AI for Dementia-Related Crash Severity Prediction: A Selective Deferral Framework to Support Human Review. Preprint at https://arxiv.org/abs/2609.22694 (2026).
65. Hendrycks, D. & Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. International Conference on Learning Representations* (2017); https://arxiv.org/abs/1610.02136.
66. Akhtyamov, R. The Harness as the Only Mutable Surface: Compliance-Bounded Self-Evolution of LLM Agents in Credit Pipelines, with a Measured Admission Gate. Preprint at https://arxiv.org/abs/2610.10629 (2026).
67. Zhou, R. et al. From Accuracy to Auditability: A Survey of Determinism in Financial AI Systems. Preprint at https://arxiv.org/abs/2605.23955 (2026).
68. Feng, D. et al. Empowering Many, Biasing a Few: Generalist Credit Scoring through Large Language Models. Preprint at https://arxiv.org/abs/2310.00566 (2023).
69. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency & Federal Deposit Insurance Corporation. *SR 26-2: Revised Guidance on Model Risk Management* (Federal Reserve, 17 April 2026); https://www.federalreserve.gov/supervisionreg/srletters/sr2602.htm.
70. Dou, Y., Lian, S. & Li, S. Conformal Cascade: Distribution-Free Accuracy Guarantees for Multi-Tier LLM Inference. Preprint at https://arxiv.org/abs/2607.25018 (2026).
71. 弘犀智能（Great Unicorn Intelligence）. SAM 数据分析与评分卡建模平台（SAM Data Analysis & Scorecard Modeling Platform）. https://www.hxallai.com (2026).
72. 弘犀智能（Great Unicorn Intelligence）. 企业智能体平台 · Harness（Enterprise Agent Platform · Harness）. https://www.hxallai.com/en/platform/agent/ (2026).

## Data availability

No real business data were used and no new experimental data were generated. Numbers quoted in the text come from the listed references and official documents; the data in the illustrative calculation are generated by simulation scripts under the settings stated in the text.

## Code availability

The scripts for the simulation and Figs. 1–3 are src/sim.py and src/figs.py (Python, fixed random seeds), provided with the article package; they reproduce every number in Table 4 and Figs. 1–3.

## Competing interests

The authors are employees of Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. The SAM data analysis and scorecard modelling platform and the Enterprise Agent Platform Harness mentioned in the text are products of the company.

Copyright Notice: © 2026 Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. All rights reserved. Any reposting, quotation in commentary, or adaptation must credit the author (Great Unicorn Intelligence Research Team) and cite the source (Great Unicorn Intelligence, https://www.hxallai.com).