KNOWING WHEN TO DEFER: TRAINING LANGUAGE MODELS TO GENERATE AN ANSWER AND ITS CONFIDENCE LEVEL TOGETHER FOR THE THRESHOLDS THAT MATTER

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model that reports its confidence level in each answer allows a user to accept the answer when that confidence level meets a threshold, and to defer it otherwise. This threshold depends on the relative costs of a deferral and of a wrong acceptance. Applications usually face only a limited range of such costs, which gives a *band* of meaningful thresholds. Previous work trains models under the Brier score, which ignores this and weights every threshold equally, so some of the training signal goes to confidence values on which no decision depends. The clipped Brier score weights only the threshold band, but cannot distinguish confidence reports on the same side outside it, and thresholds that may come into use later are not weighted either. We ask how to prioritise current decisions while accounting for learning behaviour and future adaptation. Our approach is the first built to tackle this aim. We first propose a reward that combines a binary correctness score with a weighted confidence score. Our confidence score mixes the clipped Brier score with a small full-range Brier term, the *reserve*, in a controllable proportion that we call the *reserve weight*. For a fixed answer, any positive reserve weight makes the true probability of correctness the unique minimiser of the expected score, so the band concentrates the learning signal without moving its target. We then use this score for three different training systems that each allow a confidence level to be reported for a model's generated answer: a confidence reader trained on a frozen model's features; a separate model trained to report confidence for fixed answers; and a model trained to answer and report confidence jointly by reinforcement learning. Our analysis shows how the band, the reserve weight and the score's weight against the correctness reward set the incentive to answer correctly, giving a rule for placing and weighting the band. With the answer model fixed and only the confidence trained, our score gives the best-calibrated confidence among all previous methods we evaluated, and lowers expected calibration error to less than half that of a model trained under the full-range Brier reward (0.030 against 0.067 on HotpotQA, 0.034 against 0.093 on MedQA). In joint training on MedQA, our score not only cuts *decision cost* at the targeted band against the full Brier reward, but also shows better calibration and ranking over the whole range. These tests support one theory to place the band where the model's confidence lies and weight it as strongly as the incentive to answer correctly allows, which keeps the model focused on the thresholds that matter.

1 INTRODUCTION

A large language model (LLM) can be far more useful when it can say how sure it is (Bani-Harouni et al., 2025; Wu et al., 2026). Consider a decision-maker who receives an answer together with the model's *confidence report* c, the probability it believes its own answer is correct, and who must then accept the answer or defer the question. If a wrong acceptance costs 1 and a deferral costs $d \in (0, 1)$, a decision-maker who trusts the report can minimise the *decision cost* by accepting only when $c \geq 1 - d$ and deferring otherwise. This creates a *decision threshold* $\tau = 1 - d \in (0, 1)$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

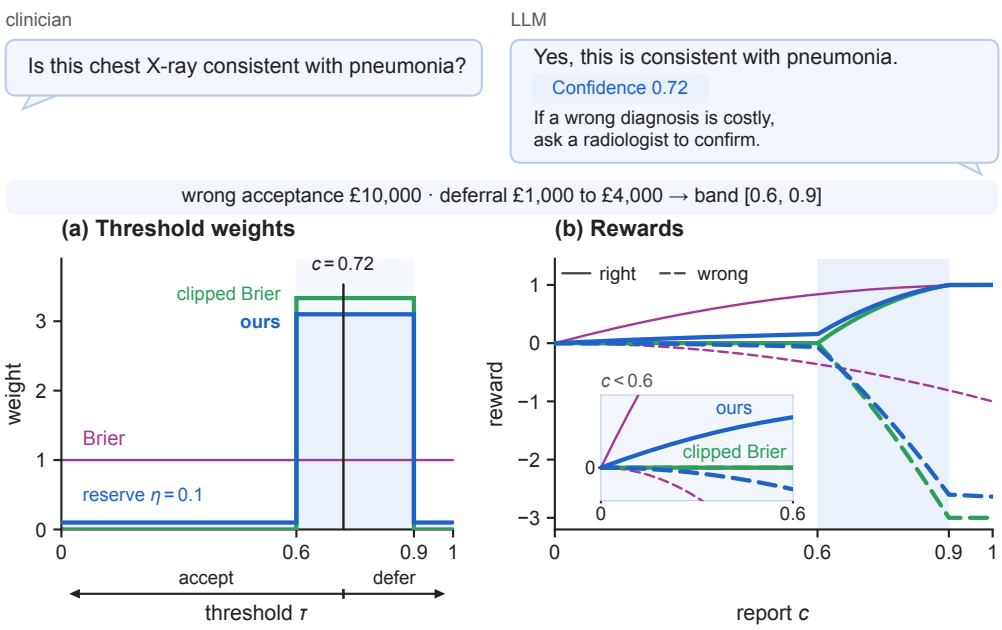


Figure 1: Left, the clinic example on the threshold axis and the weights of the Brier score (Brier, 1950), the clipped Brier score (Flores et al., 2026) and ours, with the band $[0.6, 0.9]$ shaded. Right, rewards against the confidence report for right (solid) and wrong (dashed) answers. Outside the band, the clipped Brier score ties reports 0.2 and 0.5, while ours keeps a slope.

(Chow, 1970; Elkan, 2001). As models take on more decisions, the quality of that one reported confidence number matters more.

LLMs can report such numbers through prompting or after fine-tuning (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023), yet they often hallucinate or report overconfidently, and pretraining alone cannot remove hallucination on rarely seen facts (Kalai & Vempala, 2024). The standard post-training approach, reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024), trains models on answer correctness alone, which further rewards overconfident guessing (Kalai et al., 2025; Damani et al., 2026; Wu et al., 2026). To obtain a reliable confidence report, recent work such as Reinforcement Learning with Calibration Rewards (RLCR) (Damani et al., 2026) and Wu et al. (2026) trains an LLM to answer and to report its confidence accurately at the same time. A recent model, Jev (Almeida, 2026), is also trained to produce a confidence field beside every answer.

On the other hand, in practice, the deferral cost is often known as a potential range, which gives a corresponding range of potential decision thresholds which we call the *band*. For example, consider a model making a clinical diagnosis where a wrong acceptance costs £10,000 and a deferral costs between £1,000 and £4,000. This gives the band $[0.6, 0.9]$ shown in Figure 1. However, RLCR (Damani et al., 2026) and Wu et al. (2026) both train with the *full Brier reward* (Brier, 1950), a proper scoring rule (Gneiting & Raftery, 2007) that treats every threshold τ as equally likely (Buja et al., 2005; Flores et al., 2026). In our example, this means 70% of that reward's weight goes to deferral costs below £1,000 or above £4,000, which the clinic never faces, so training and deployment are misaligned. To tackle this misalignment, Flores et al. (2026) propose an evaluation metric, the *clipped Brier score*, which assigns weight only to the thresholds inside the band. However, outside the band it cannot tell one report from another on the same side: a report of 0.2 scores exactly as 0.5 does, even though the two may lead to different decisions if the band moves. The problem, then, is how to prioritise current decisions while still accounting for learning behaviour and future adaptation.

Our *weighted score* adds a full-range Brier *reserve*, scaled by a small *reserve weight*, to the clipped Brier score for deployment bands. We develop it for frozen-model readers, separate reporters and joint answer–confidence training, with the following key contributions:

Table 1: The costs of a decision. Left, the cost of each action on a right and on a wrong answer, in units of a wrong acceptance: accepting an answer that is right with probability p costs $1 - p$ in expectation and deferring costs d, so accepting costs no more than deferring exactly when $p \geq 1 - d = \tau$. Right, the regret of Eq. 1: the same cost minus that of a decision-maker who accepts every right answer and defers every wrong one, written with $\tau = 1 - d$.

cost	right (p)	wrong ($1 - p$)
accept	0	1
defer	d	d

regret	right (p)	wrong ($1 - p$)
accept	0	τ
defer	$1 - \tau$	0

- We introduce the first LLM confidence score for bands with a reserve and its first use in training. For fixed answers, a positive reserve uniquely favours true probabilities in expectation (Schervish, 1989; Gneiting & Raftery, 2007). We derive the largest score weight that preserves correctness incentives and split decision cost into band regret and answer error (Sections 3 and 4.1).
- With the answer model fixed and only the confidence trained, the *band reader* is a classifier fitted in minutes on frozen features. On HotpotQA (Yang et al., 2018), it beats classical readers and RLCR's full-parameter joint training on every reported calibration and ranking mean. A separate *band reporter*, trained only on the confidence report, leads every classical reader and reporter and RLCR on MedQA (Jin et al., 2020) in Brier score and calibration error, and ranks within 0.001 of the best classical reporter. Their calibration errors are 0.030 against 0.067 and 0.034 against 0.093, respectively, each under half of RLCR's. The band reporter's error is also the lowest among evaluated MedQA methods. Against RLCR, HotpotQA's band readers lower mean decision cost at their training bands, and the middle reader does so on all bands, whilst MedQA's band reporter lowers it at the high band (Section 4.2).
- MedQA's jointly trained middle and high arms beat RLCR's mean decision cost at their respective training bands, with the high band leads jointly trained models in calibration and ranking at a small accuracy cost. This suggests reduced pressure towards extremes outside the band (Section 4.3).
- Our rule, derived from the reward gap and confirmed on HotpotQA, places the training band where confidence lies and weights it as far as the answer's learning signal allows. The rule supports adaptation during training. An ablation with the *equal-weight setting*, which weights thresholds inside the band as the Brier score does (Section 3), shows the trade-off (Section 4.4).

2 PRELIMINARIES

We write x for a question, $\hat{a}$ for the model's answer, c for its confidence report and $y \in \{0, 1\}$ for the answer's correctness, with $y = 1$ when the answer is right. Our goal is to train the model to produce c that aligns with the true epistemic probability p, which Bani-Harouni et al. (2025) describe as "the actual likelihood of correctness given the model's internal knowledge state". The report is *calibrated* when $c = p$ (Gneiting & Raftery, 2007; Guo et al., 2017). As Table 1 shows, a calibrated report $c = p$ makes $\tau = 1 - d$ the best decision threshold. When the deferral cost d is known, a user can therefore minimise their expected decision cost by accepting when $c \geq \tau$ and deferring otherwise (Flores et al., 2026).

For a given threshold τ, *regret* $r_\tau(c, y)$ (Flores et al., 2026) is the extra decision cost of acting on c at τ, compared with a model that accepts every right answer and defers every wrong one:

$$r_\tau(c, y) = \tau\, \mathbf{1}[y = 0,\ c \geq \tau] + (1 - \tau)\, \mathbf{1}[y = 1,\ c < \tau]. \tag{1}$$

If the threshold is uniformly distributed on $[0, 1]$, the average regret is half the Brier score (Buja et al., 2005; Hernández-Orallo et al., 2012; Gneiting & Raftery, 2007):

$$(c - y)^2 = 2 \int_0^1 r_\tau(c, y)\, \mathrm{d}\tau. \tag{2}$$

This score is strictly proper (Schervish, 1989; Buja et al., 2005; Gneiting & Raftery, 2007), so for a fixed answer the calibrated confidence $c = p$ is the unique minimiser of the expected score. More generally, every proper score is an average of regret over a distribution of thresholds (Schervish, 1989; Buja et al., 2005); the Brier score is the uniform case, and the scores below change that distribution. Previous work, RLCR (Damani et al., 2026) and Wu et al. (2026), defines its reward as the correctness signal minus the Brier score, $y - (c - y)^2$, which we call the *full Brier reward*. By Eq. 2, this reward treats thresholds as equally likely on $[0, 1]$, whatever range the deferral cost covers.

For a band $[a, b] \subsetneq [0, 1]$ with $a < b$, the clipped Brier score is defined as $\frac{2}{b-a} \int_a^b r_\tau(c, y)\, \mathrm{d}\tau$. Flores et al. (2026) designed it to evaluate a model only over the band of thresholds associated with the range of potential deferral costs. When used directly for training, the clipped Brier score is constant over reports below the band and constant over reports above it.

Appendix A gives the proofs and details of this section.

3 Weighting the thresholds that matter

We want learning to concentrate on the band's decisions while still distinguishing reports elsewhere. The weighted score serves as the band reader's loss on a frozen model. Multiplied by a positive weight, it supplies the confidence term for training the band reporter, a separate model reporting confidence in a fixed answer. It serves the same role when training the answer and its confidence together.

We combine the clipped Brier score (Flores et al., 2026) with the Brier score (Brier, 1950), using reserve weight $\eta \in [0, 1]$ and weight $\lambda > 0$ on the weighted score:

$$R(c, y) = y - \lambda S_\eta(c, y),$$
$$S_\eta(c, y) = (1 - \eta) \underbrace{\frac{2}{b - a} \int_a^b r_\tau(c, y)\, \mathrm{d}\tau}_{\text{clipped Brier score}} + \eta \underbrace{(c - y)^2}_{\text{Brier score}}. \tag{3}$$

In the classical threshold-weight representation, its threshold weight is a probability density with value $\eta + (1 - \eta)/(b - a)$ inside the band and η outside (Schervish, 1989; Buja et al., 2005).

With unit weight λ, we recover the full Brier reward when either $\eta = 1$ or the band is $[0, 1]$ (Damani et al., 2026; Wu et al., 2026). At $\eta = 0$, S_η is exactly the clipped Brier score. With no reserve, shrinking the band to a single threshold recovers the point member for reports distinct from that threshold. Its contribution $-\lambda S_\eta$ to the reward recovers the calibration term of C3RL (Yang et al., 2026) and the reliability term of Deliberative Searcher (Yin et al., 2026) at their threshold, up to positive affine maps and C3RL's tie convention. Our contribution within the classical family is the narrower band with a positive reserve, used to train confidence alone or together with the answer.

Two properties make our score safe to train with: the largest score weight that still rewards a right answer at least as much as a wrong one at the same report (Proposition 1); and honesty, since any positive reserve makes the true probability the unique best report (Proposition 2).

Proposition 1 (the correctness incentive). For $\eta > 0$, a right answer earns at least as much as a wrong one at every common report $c \in [0, 1]$ if and only if $\lambda \leq 1/A$, where $A = \eta + (1-\eta)(2-a-b)$. Within this bound, the two rewards are equal only at report zero when $\lambda = 1/A$. At that limit, the reward gap below the band is $R(c, 1) - R(c, 0) = 2\lambda\eta c$ for $c \leq a$. This preserves the correctness incentive (Damani et al., 2026; Wu et al., 2026).

The gap below the band in Proposition 1 makes the role of the reserve explicit: the correctness incentive there rests on the reserve alone. The gap is positive for every positive report below the band and narrows towards zero as the report falls.

We choose the limit setting because it gives the weighted score the greatest weight compatible with the correctness incentive. Section 4 trains with $\eta = 0.1$. The corresponding limits are $\lambda = 20/29$ at the low band $[0.1, 0.4]$, 1 at the middle band $[0.3, 0.7]$ and $20/11$ at the high band $[0.6, 0.9]$.

Proposition 2 (honesty). For a fixed answer with correctness probability p, any positive reserve weight makes $c = p$ the unique minimiser of the expected weighted score over $c \in [0, 1]$. The

score is strictly proper because its threshold weight is at least $\eta > 0$ everywhere (Buja et al., 2005; Gneiting & Raftery, 2007). For $\lambda > 0$, the same report uniquely maximises the expected reward, preserving the calibration incentive (Damani et al., 2026; Wu et al., 2026). This holds whether p is inside or outside the band.

At the limit setting, most threshold weight lies inside the band and positive weight remains outside it.

The equal-weight setting in Section 4.4 chooses λ and η so that λS_η gives thresholds inside the band the same weight as the Brier score.

The distinction between the clipped term and the reserve is explicit in the closed form. Write $h = b - a$ for the band's width, $z = \min\{\max\{c, a\}, b\}$ for the report clipped to the band, and $o_y = b$ when $y = 1$ and $o_y = a$ when $y = 0$. Then

$$2 \int_a^b r_\tau(c, y) \, \mathrm{d}\tau = (y - z)^2 - (y - o_y)^2. \tag{4}$$

Dividing by h gives the clipped Brier score in Eq. 3. Its dependence on c is through z, which stays fixed as a report moves below a or above b. The reserve uses c itself and removes the clipped score's plateaus (Flores et al., 2026).

The equal-weight setting weights the clipped quadratic rule of Flores et al. (2026, Proposition D.4) by 0.9 and the Brier score by 0.1. The clipped rule is the right-hand side of Eq. 4, or h times the clipped Brier score. In Eq. 3, this sets $\lambda = 0.1 + 0.9h$ and $\eta = 0.1/\lambda$. The ablation therefore changes both the weight on the weighted score and the reserve weight.

Eq. 3 applies to any policy returning an answer and a report, with any reinforcement-learning method. Section 4 tests the score in its three uses. Other threshold weights, including the classical Beta and Gaussian families, remain future work (Buja et al., 2005). Peaked Beta weights have already been used in LLM forecasting (Turtel et al., 2026), and Wu et al. (2026) use a truncated Beta weight. Appendix A gives the proofs, Appendix B the maps between reward settings, and Figure 4 the reward's shape.

4 Experiments

4.1 Setup and the evaluation

We use decision cost to distinguish gains in confidence from accuracy gains when evaluating the score's three uses. Including baselines, the comparison follows four paths on HotpotQA (multi-hop questions) and MedQA (medical-exam questions) using Qwen3-4B-Instruct-2507 (Qwen Team, 2025). Path A reads untrained confidence from answer probability (Kadavath et al., 2022) or verbalised reports. Path B follows RLCR's classical baselines (Damani et al., 2026), with three readers and one trained reporter. The readers are a linear probe fitted with binary cross-entropy (BCE), a Brier classifier and recalibration, and the reporter uses the full Brier reward. For each band, Path C fits a band reader or trains a band reporter on the classical readers' correctness-only checkpoints. Path D trains answers and confidence jointly under the limit setting, RLCR or the equal-weight setting. Joint training uses group relative policy optimisation (GRPO) (Shao et al., 2024) with Dr. GRPO (Liu et al., 2025), and separate reporters use low-rank adaptation (LoRA) (Hu et al., 2022) on the report step alone.

Within each task and seed slot, the probe and classifiers in Paths B and C use saved features from the same frozen correctness-only checkpoint. They use the same $2{,}000$ fitting questions, disjoint from the held-out set. Every fitted system is scored on the same held-out set for its task, $1{,}000$ questions for HotpotQA and the $1{,}273$-question MedQA test set. The LoRA reporters train on the same stored answers, one per fitting question, generated by the frozen checkpoint at the evaluation cap. The matched full-Brier reporter uses RLCR's reward on those same stored answers. This gives Paths B and C in Table 2 a shared data protocol for comparing BCE, the Brier score, the full Brier reward and the weighted score. For the weighted score, we fit band readers and train band reporters at both the limit and equal-weight settings defined in Section 3.

We evaluate decision cost at the low $[0.1, 0.4]$, middle $[0.3, 0.7]$ and high $[0.6, 0.9]$ bands and over $[0, 1]$. The cost identity splits it into band regret (the confidence's share) and the answers' error

Table 2: Decision cost, calibration and ranking across the four paths. The band reader on HotpotQA and band reporter on MedQA achieve the lowest ECE without retraining the answer model. Composite rows use each band's own reader or reporter for band costs and the middle one's Brier, ECE, AUROC and accuracy. Bold marks the best value in each column within each task; shaded rows are our systems, the limit setting darker and the equal-weight setting lighter. Cost$[0, 1] = (\text{Brier} + 1 - \text{accuracy})/2$.

	decision cost ↓						
path system	low	middle	high	Brier ↓	ECE ↓	AUROC ↑	accuracy ↑
HotpotQA, the frozen model's accuracy .554							
A answer probability	.445	.429	.292	.273	.269	.806	—
A verbalised report	.443	.426	.377	.348	.367	.755	—
B linear probe / BCE classifier	.409	.327	.207	.157	.073	.854	—
B Brier classifier	.410	.327	.206	.156	.065	.851	—
B recalibration	.432	.375	.236	.209	.085	.755	—
B full-Brier reporter	.428	.346	.224	.185	.068	.795	.555
C band reader (ours)	**.406**	**.324**	**.202**	.149	**.030**	**.859**	—
C band reporter (ours)	.430	.348	.225	.187	.079	.793	.556
D RLCR (full Brier reward)	.410	.333	.209	.168	.067	.825	**.562**
D ours, low band	.416	.359	.239	.190	.132	.807	.548
D ours, middle band	.422	.336	.213	**.147**	.053	.857	.527
D ours, high band	.482	.373	.220	.157	.073	.852	.467
D equal-weight setting, low band	.434	.383	.243	.216	.076	.680	.556
D equal-weight setting, middle band	.435	.361	.235	.205	.144	.779	.548
D equal-weight setting, high band	.453	.374	.219	.210	.114	.754	.555
MedQA, the frozen model's accuracy .757							
A answer probability	.243	.244	.207	.183	.125	.725	—
A verbalised report	.243	.241	.223	.205	.184	.682	—
B linear probe / BCE classifier	.250	.237	.175	.167	.062	.734	—
B Brier classifier	.248	.238	.179	.168	.056	.723	—
B recalibration	.245	.237	.190	.173	.068	.682	—
B full-Brier reporter	.243	.232	.174	.160	.051	.750	.758
C band reader (ours)	.244	.245	.179	.185	.142	.725	—
C band reporter (ours)	.243	.233	.168	.159	**.034**	.749	.757
D RLCR (full Brier reward)	.229	.224	.183	.164	.093	.730	.771
D ours, low band	.240	.237	.215	.200	.161	.641	.760
D ours, middle band	**.218**	**.213**	.184	.164	.095	.705	**.779**
D ours, high band	.237	.216	**.165**	**.136**	.069	**.810**	.746
D equal-weight setting, low band	.247	.304	.237	.234	.174	.559	.753
D equal-weight setting, middle band	.233	.231	.205	.192	.116	.603	.767
D equal-weight setting, high band	.245	.244	.193	.176	.083	.712	.756

$(1 - \text{accuracy})(1 - m)$ for band midpoint m. Brier measures probabilistic accuracy, expected calibration error (ECE) (Naeini et al., 2015; Guo et al., 2017) diagnoses calibration, and the area under the receiver operating characteristic curve (AUROC) diagnoses ranking. Every system uses three seeds. Paired comparisons match questions within seed slots and report bootstrap intervals. We report the cut share, the fraction of answers cut at the length cap.

4.2 Better confidence without retraining the model

The band reader on HotpotQA and the band reporter on MedQA beat RLCR's full-parameter joint training on Brier, ECE and AUROC and have each task's lowest ECE among the systems in Table 2. Their ECEs are 0.030 against 0.067 and 0.034 against 0.093, respectively, each under half RLCR's. On HotpotQA, the reader also beats every classical reader on these diagnostics and lowers mean decision cost at every training band. A single middle-band reader lowers mean decision cost against every classical reader and RLCR across all evaluation bands and the whole range. Its whole-range cost is 0.297 against the best classical reader's 0.301. Table 3 gives every HotpotQA reader's mean cost at every evaluation band, each fitted band shown separately, so it holds the fixed-band comparison of losses under one fitting protocol. On HotpotQA, at each band the band reader fitted for it beats

Reader	Fitted band	Low	Middle	High	$[0, 1]$
Linear probe / BCE classifier	–	0.409	0.327	0.207	0.301
Brier classifier	–	0.410	0.327	0.206	0.301
Recalibration	–	0.432	0.375	0.236	0.327
Band reader, limit setting	Low	0.406	0.330	0.212	0.302
Band reader, limit setting	Middle	0.406	0.324	0.202	0.297
Band reader, limit setting	High	0.408	0.322	0.202	0.298
Band reader, equal-weight	Low	0.406	0.327	0.208	0.300
Band reader, equal-weight	Middle	0.406	0.324	0.202	0.297
Band reader, equal-weight	High	0.408	0.322	0.200	0.297

Table 3: HotpotQA mean decision cost at every evaluation band for each reader fitted under the protocol of Section 4.1, three seeds, the band readers by fitted band; lower is better.

First method minus second	Evaluation band	Cost difference	95% interval
Limit setting, high − RLCR	High	-0.018	$[-0.026, -0.010]$
Limit setting, middle − RLCR	Middle	-0.011	$[-0.022, +0.000]$
Limit setting, high − limit setting, middle	High	-0.019	$[-0.026, -0.011]$
Limit setting, high − limit setting, middle	Middle	$+0.003$	$[-0.007, +0.014]$
Band reporter, limit setting, high − full-Brier reporter	High	-0.006	$[-0.009, -0.004]$
Band reporter, equal-weight, high − RLCR	High	-0.013	$[-0.019, -0.007]$

Table 4: Paired MedQA cost differences over three seeds and 1,273 shared questions, with 95% bootstrap intervals from 10,000 resamples. Negative differences favour the first method; the high arm's high-band gains exclude zero.

every classical reader: 0.406, 0.324 and 0.202 against the linear probe's 0.409, 0.327 and 0.207. By the cost identity in Section 4.1, the reader's confidence gain outweighs RLCR's small accuracy advantage at every band. On MedQA, the band reporter leads the frozen-answer systems on Brier and ECE, with AUROC close to the full-Brier reporter's, and lowers high-band cost against that reporter by 0.006, with a 95% interval of $[0.004, 0.009]$. At the high band, it has the lowest mean decision cost among the reported frozen-answer systems, 0.168 against the next lowest 0.170, and also beats RLCR's 0.183. Reader–RLCR comparisons by band and limit-setting reporter comparisons with RLCR or the equal-weight reporter use means without paired intervals.

The linear probe uses the same features and answers as the HotpotQA band reader. Both settings' HotpotQA readers beat the linear probe's mean own-band cost, with at most 0.002 between them. Only the reserve rewards reports beyond b for right answers or below a for wrong answers. Our reading is that this leaves spare capacity to order answers and set confidence inside the band, explaining the whole-range gains. HotpotQA favours the reader and MedQA the reporter when answers are fixed, while joint training changes both answer and confidence.

4.3 Also when trained together with the answer

Joint training also improves decisions where the band covers the model's confidence: MedQA's middle and high arms beat RLCR's mean cost at their respective training bands, and the high arm wins outright. Its cost difference from RLCR at the high band is -0.018, with a 95% interval of $[-0.026, -0.010]$ (Table 4). Its mean Brier score of 0.136 and AUROC of 0.810 are the best among all MedQA systems in Table 2. It also has the lowest ECE among jointly trained models, at a small cost in accuracy of 2.5 percentage points. The cut shares track truncation alongside accuracy: the high arm's per-seed values are 0.028, 0.038 and 0.129, against RLCR's 0.003, 0.001 and 0.001. The middle arm leads on mean cost at the low and middle bands and over the whole range, with no loss of accuracy. Its middle-band difference from RLCR is -0.011, with interval $[-0.022, +0.000]$, so it leads on the mean with the interval touching zero. This gain in decision cost accompanies slightly higher ECE and lower AUROC than RLCR's.

In Figure 2, the high arm beats the middle arm at the high band by 0.019, with interval $[0.011, 0.026]$ (Table 4). At the middle band it trails on the mean, with an interval including zero. At the high band,

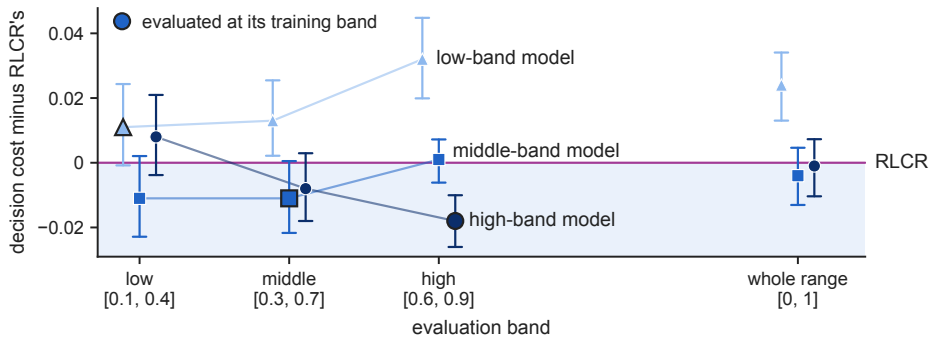


Figure 2: MedQA decision cost relative to RLCR for the three limit-setting training bands. Outlined markers denote evaluation at the training band. The middle and high arms exchange the lowest mean cost between their bands. Bars show 95% paired bootstrap intervals over the 1,273 test questions and three paired seeds (10,000 resamples).

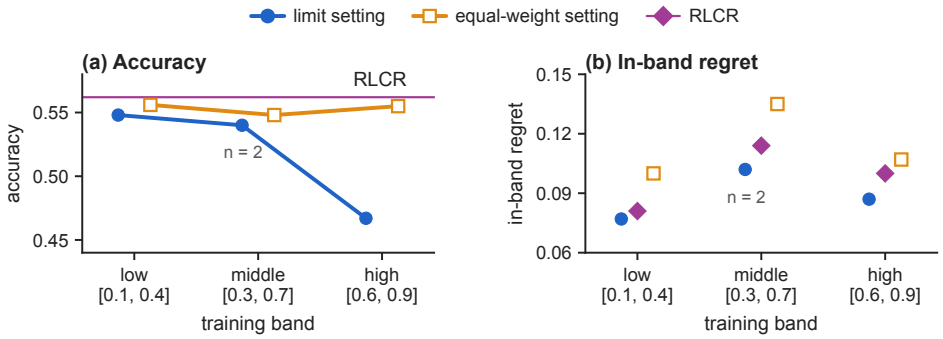


Figure 3: HotpotQA accuracy and regret at each arm's training band. The high-band limit setting improves confidence at a cost in accuracy, while the equal-weight setting retains accuracy. Middle-band limit-setting markers show the two-seed subset; Table 2 gives the three-seed means.

Section 4.1's cost identity shows that the confidence gain outweighs the high arm's accuracy cost against both RLCR and the middle arm. Section 4.2's interpretation also applies when answers and confidence share the learning signal. On HotpotQA, the high band instead lies above confidence, making the answer's signal decisive.

4.4 Where the learning signal goes: placing and weighting the band

The reward gap in Section 3 and the experiments give a practical rule for the training band, which may differ from the deployment band that the costs fix: place it where the model's confidence lies, and weight it as far as the answer's signal allows. Raising the training band on HotpotQA under the limit setting lowers accuracy from 0.548 through 0.527 to 0.467. Confidence improves at the high band, with band regret 0.087 against RLCR's 0.100. This trade-off is consistent with the below-band correctness gap in Section 3. The recorded signal measurements assign 68–79% of the high arms' advantage variance, the variance of group-centred rewards, to confidence, against about 30% under RLCR. This mechanism also offers an explanation for the MedQA high arm's smaller accuracy cost. MedQA's low arm illustrates the complementary placement mismatch, in which a band far below the reports leaves little improvement in confidence.

Figure 3 shows the equal-weight setting at the other end of the trade-off. At HotpotQA's high band it retains accuracy at 0.555, with band regret 0.107, giving up the limit setting's confidence gain. On MedQA, its high arm still improves ECE over RLCR, at 0.083 against 0.093. The frozen-answer benefit also persists, with this setting's MedQA high-band reporter beating RLCR by 0.013 in cost

with interval $[0.007, 0.019]$. These are two settings of one score, whose parameters change together, they illustrate the accuracy–confidence trade-off; adapting placement and weight during training remains future work. Appendix C gives the recipe, the gates and the compute with the evaluation protocol, and Appendix D the full per-band, per-seed and equal-weight tables.

5 Related work

How a proper score weights decision thresholds is a classical design choice, with precedents for peaking the weight at costs of interest and combining full-range and restricted scores (Schervish, 1989; Buja et al., 2005; Gneiting & Raftery, 2007; Merkle & Steyvers, 2013; Wessel et al., 2025). The clipped Brier score restricts that weight to a band for evaluation (Flores et al., 2026). Our decision-cost evaluation belongs to the tradition of judging confidence by decision value, alongside selective prediction's error rate and share of retained answers, decision curves' net benefit, forecast savings and loss averaged over operating conditions (Geifman & El-Yaniv, 2017; Vickers & Elkin, 2006; Murphy, 1977; Hernández-Orallo et al., 2012). For this evaluation, paths A and B of Section 4 compare untrained confidence, fitted readers, recalibration and a full-Brier reporter at every band with answers fixed (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Guo et al., 2017). A model trained by reinforcement learning can report overconfidently (Bereket & Leskovec, 2025). Path C's band reader and band reporter learn under our score, following work that scores confidence alone. That work uses a clipped logarithmic score for a separate report (Bani-Harouni et al., 2025) or a quadratic reward for jointly sampled answers, rationales and confidence (Xu et al., 2024). LLM forecasting uses proper terminal rewards, including those with peaked Beta weights, and outcomes differ across threshold weights (Turtel et al., 2026). Among the closest joint recipes, RLCR, the nearest recipe, and Wu et al.'s full Brier reward weight every threshold equally, while Wu et al.'s truncated logarithmic weight favours the extremes (Damani et al., 2026; Wu et al., 2026). C3RL and Deliberative Searcher's confidence terms weight one threshold and correspond to the point member of Section 3 (Yang et al., 2026; Yin et al., 2026). RLSR rewards confidence rankings for selective prediction and leaves open training for a known target risk, the error rate allowed among retained answers (Luo et al., 2026). The family, the clipped Brier score, the separate confidence step and the point member come from this prior work. We add the band with a reserve as the threshold prior for training: no prior work trains a language model's confidence, alone or with its answer, under a score weighted towards a band of decision thresholds chosen from a deployment's costs with a reserve on every other threshold.

6 Conclusion and future work

We introduced the first weighted score with a reserve for training LLM confidence at deployment bands and evaluated decision cost at those bands using one model on two tasks. With the answer model fixed and only the confidence trained, HotpotQA's band reader and MedQA's band reporter had the lowest ECE among evaluated methods, beating classical readers on whole-range calibration and RLCR's full-parameter joint training on calibration and ranking. In joint training on MedQA, the high-band arm cut decision cost at its band outright, with the best Brier score and AUROC of every system on that task, and the middle-band arm beat RLCR on mean cost at its band while keeping accuracy. Our rule, derived from the reward gap and confirmed on HotpotQA, places the training band where confidence lies and weights it as far as the answer's signal allows. Future work includes adapting the band's placement and weight to the model's confidence during training and testing larger models on more tasks. We could also explore smooth peaked weights (Buja et al., 2005; Turtel et al., 2026) and narrower bands with the reserve in the same construction, lengthen confidence reports, report confidence for answer chunks and train with the score for a known target risk (Luo et al., 2026).

Reproducibility statement

Appendix C gives everything needed to reproduce our results: the model, the datasets and their splits, the reward settings with their coefficients, the GRPO and LoRA hyperparameters, the training gates, the reader-fitting protocol, the seeds and the paired bootstrap procedure; Appendix A gives the full proofs and Appendix D the per-band and per-seed results behind every number in Table 2 and the

text. Code, the exact run table, prompts, seeds and the evaluation scripts will be released together with the camera-ready version.

THE USE OF LARGE LANGUAGE MODELS

We built this work so that every statement in the paper is traceable to evidence. Large language model assistants were all run through a harness the authors built from scratch, and supported the process by assisting the writing and polishing, by helping to implement and test the training and analysis code against verifiable local data, and by gathering related work for the authors to read. They also helped check the paper before the final version was submitted, in blind passes by independent model instances that had no part in producing it but had access to the same verifiable local data: each proof step in Appendix A against its stated assumptions, each number in the text and tables against the run record it comes from, and each citation against the cited source in local storage. The authors reviewed all of it last. The research questions, the score, the theory, the experimental design and every claim are the authors' own, and the authors take full responsibility for the content.

REFERENCES

Diogo Almeida. Introducing system one models & jev. TypeSafe AI blog, 2026. 15 September 2026.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. *arXiv preprint arXiv:2503.02623*, 2025.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: GRPO induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, Wharton School, University of Pennsylvania, and University of Washington, 2005. Manuscript dated 3 November 2005.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.16806.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Gerardo Flores, Abigail Schiff, Alyssa H. Smith, Julia A. Fukuyama, and Ashia C. Wilson. A consequentialist critique of binary classification evaluation: Theory, practice, and tools. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026. arXiv:2504.04528.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13(91):2813–2869, 2012.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 160–171, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025.

Gaoxiang Luo, Yifan Wu, Sinian Zhang, Aryan Deshwal, and Ju Sun. Aligning language models with selective prediction. *arXiv preprint arXiv:2607.03528*, 2026.

Edgar C. Merkle and Mark Steyvers. Choosing a strictly proper scoring rule. *Decision Analysis*, 10 (4):292–304, 2013. doi: 10.1287/deca.2013.0280.

Allan H. Murphy. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105:803–816, 1977.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL `https://arxiv.org/abs/2505.09388`.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989. doi: 10.1214/aos/1176347398.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Emir H. Shuford, Arthur Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. doi: 10.1007/BF02289503.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Benjamin Turtel, Paul Wilczewski, Kris Skotheim, Ville A. Satopää, and Philip E. Tetlock. How proper scoring rules shape LLM forecasting. *arXiv preprint arXiv:2608.28482*, 2026.

Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

Jakob Benjamin Wessel, Maybritt Schillinger, Frank Kwasniok, and Sam Allen. Enforcing tail calibration when training probabilistic forecast models. *arXiv preprint arXiv:2506.13687*, 2025.

Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. Mitigating llm hallucination via behaviorally calibrated rl. *arXiv preprint arXiv:2512.19920*, 2026.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales. *arXiv preprint arXiv:2405.20974*, 2024.

Xuqing Yang, Yi Yuan, Shanzhe Lei, and Xuhong Wang. Scaling with confidence: Calibrating confidence of llms for adaptive test time scaling. *arXiv preprint arXiv:2607.01612*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Zhenyun Yin, Shujie Wang, Xuhong Wang, Xingjun Ma, and Yinchun Wang. Deliberative searcher: Improving llm reliability via reinforcement learning with constraints. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4340–4354, 2026.

A PROOFS

Section 2 uses $c = p$, where $p = \Pr(y = 1 \mid \mathcal{I})$ is the correctness probability given the model's internal knowledge state for a fixed answer. Reporting p implies standard calibration, $\Pr(y = 1 \mid c) = c$ (Guo et al., 2017; Naeini et al., 2015), but calibration alone need not identify p. Indeed, if c depends only on $\mathcal{I}$ and randomisation independent of y given $\mathcal{I}$, then $\mathbb{E}[y \mid c] = \mathbb{E}[p \mid c]$.

The cost in Table 1 is $r_\tau(c, y) + (1 - y)(1 - \tau)$, the regret plus the oracle's cost. Writing $h = b - a$ and $m = (a + b)/2$, averaging over answers and thresholds uniform on the band gives

$$\text{decision cost} = \underbrace{\frac{1}{h}\int_a^b \mathbb{E}\, r_\tau(c, y)\, \mathrm{d}\tau}_{\text{band regret}} + (1 - \text{accuracy})(1 - m).$$

Only the band regret depends on the reports, so at fixed accuracy minimising it minimises the decision cost.

The Brier identity in Eq. 2 and the clipped form below follow from Flores et al. (2026, Theorems 2.4 and 3.3), taking their positive class as a wrong answer, their positive prediction as a deferral and their cost ratio as $1 - \tau$. Both scores equal twice their respective average regret, a factor omitted in those theorem statements but present in the final line of their proof of Theorem D.5. Flores et al. credit the Brier identity to Shuford et al. (1966). Flores et al.'s tie convention differs at $c = \tau$, which does not affect these integrals. With $z = \min\{\max\{c, a\}, b\}$ and $o_y = b$ for $y = 1$ and a for $y = 0$, the closed form is Eq. 4 in Section 3; dividing by h gives the clipped Brier score of Section 2.

In the classical threshold-weight representation (Schervish, 1989; Buja et al., 2005), substituting Eq. 2 into Eq. 3 gives

$$S_\eta(c, y) = 2\int_0^1 w(\tau)\, r_\tau(c, y)\, \mathrm{d}\tau, \qquad w(\tau) = \eta + \frac{1 - \eta}{h}\, \mathbf{1}[a \le \tau \le b].$$

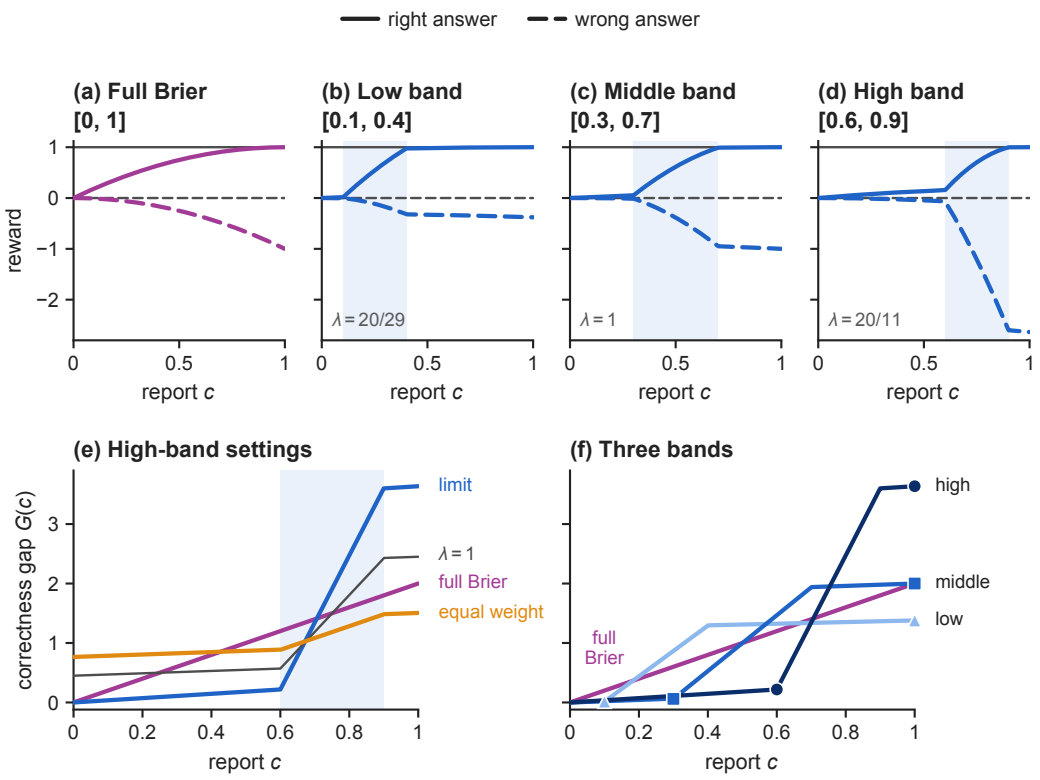


Figure 4: Rewards for right (solid) and wrong (dashed) answers at the three bands (top), with the full Brier reward (Damani et al., 2026) and correctness-only reward for comparison, and correctness gaps (bottom). At the limit setting, $\eta = 0.1$ and $\lambda = 1/A$, the gap below the band is $2\lambda\eta c$, so the correctness incentive rests on the reserve alone.

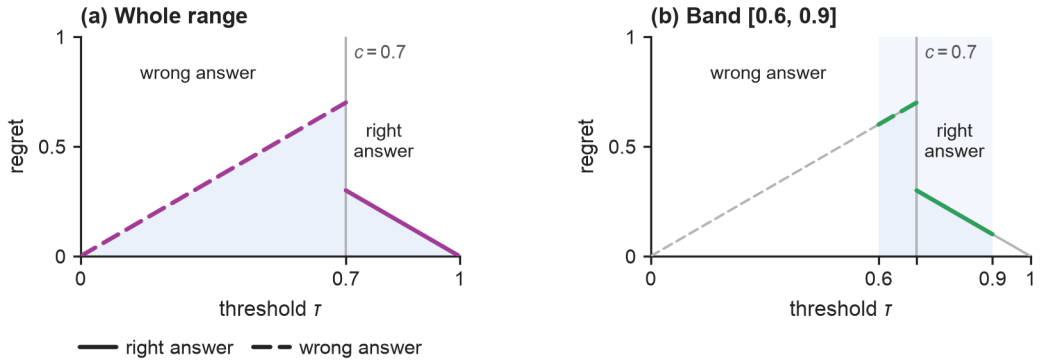


Figure 5: A score as an area over the threshold axis. Left: for a report $c = 0.7$ the regret of accepting a wrong answer (τ, dashed) and of deferring a right one ($1-\tau$, solid) integrate to $\frac{1}{2}c^2$ and $\frac{1}{2}(1-c)^2$, respectively half the Brier score for a wrong answer and half the Brier score for a right answer (Brier, 1950). Right: the clipped Brier score (Flores et al., 2026) is twice the corresponding regret area restricted to $[0.6, 0.9]$, divided by the band's width 0.3.

The weight integrates to one, with share $1-\eta+\eta h$ on the band and $\eta(1-h)$ outside. For Section 4's limit setting, $\eta = 0.1$; multiplying by $\lambda = 1/A$ leaves these shares unchanged.

For a fixed answer with $y \sim \mathrm{Bernoulli}(p)$, the excess formula of Buja et al. (2005, Section 20) gives

$$\mathbb{E}\, S_\eta(c, y) - \mathbb{E}\, S_\eta(p, y) = 2 \int_{\min(c,p)}^{\max(c,p)} w(\tau)\, |\tau - p|\, \mathrm{d}\tau \geq \eta (c - p)^2. \tag{5}$$

Any positive reserve charges every open interval, the strict-propriety condition of Gneiting & Raftery (2007, Theorem 3), so p uniquely minimises the expected score and maximises the expected reward for $\lambda > 0$. At $\eta = 0$, reports in $[0, a]$ tie when $p < a$, and reports in $[b, 1]$ tie when $p > b$ (Flores et al., 2026, Proposition D.4); the positive reserve removes these plateaus.

For the correctness gap, subtracting the outcomes in Eq. 3 using Eq. 4 gives

$$D(c) = S_\eta(c, 1) - S_\eta(c, 0) = A - 2\left[\eta c + \frac{1-\eta}{h}(z - a)\right],$$

$$G(c) = R(c, 1) - R(c, 0) = 1 - \lambda A + 2\lambda\left[\eta c + \frac{1-\eta}{h}(z - a)\right],$$

where $A = \eta + (1 - \eta)(2 - a - b) > 0$ for a positive reserve. Since z is non-decreasing in c, a positive reserve makes G strictly increasing, with minimum $G(0) = 1 - \lambda A$. Hence $G(c) \geq 0$ for every report if and only if $\lambda \leq 1/A$; at the limit $\lambda = 1/A$, equality occurs only at $c = 0$. Below the band, $z = a$, so at the limit $G(c) = 2\lambda\eta c$.

With our coefficient convention, Damani et al. (2026, Lemma 2) give the correctness condition $\lambda[S_\eta(p, 1) - S_\eta(p, 0)] \leq 1$ for every p, ensuring that the best expected reward of an answer does not decrease with its probability of being right. This is exactly $G(p) \geq 0$, so the bound above preserves the correctness incentive together with strict propriety, including at the limit setting.

B THE SCORE'S DERIVATION

The coefficient limit $\lambda = 1/A$ with $A = \eta + (1 - \eta)(2 - a - b)$, its values at $\eta = 0.1$, and the equal-weight setting's coefficients $\lambda = 0.1 + 0.9h$ and $\eta = 0.1/\lambda$ are stated in Section 3, with the derivation of the limit in Appendix A.

With $\eta = 0$, the score tends to $2r_{\tau_0}(c, y)$ as the band shrinks to τ_0 whenever $c \neq \tau_0$. At $c = \tau_0$, the point member uses the accepting tie of Eq. 1; this is an atomic threshold measure in the classical representation (Schervish, 1989; Buja et al., 2005). At $\tau_0 = 1/2$ and $\lambda = 1$, Deliberative Searcher's reliability term (Yin et al., 2026) is $1 - 2r_{1/2}(k/10, y)$ for its report $k \in \{0, \ldots, 10\}$. This shifts the point member's contribution $-\lambda S_\eta$ to the reward by one. C3RL's calibration term (Yang et al., 2026) is $\beta\big(1 - 4r_{1/2}(k/10, y)\big)$ for $k \in \{1, \ldots, 10\}$ with $k \neq 5$ and $\beta > 0$, which scales that contribution by 2β and shifts it by β. At $k = 5$, C3RL instead assigns $\beta(1 - 2y)$, whereas $\beta(1 - 4r_{1/2}(1/2, y)) = \beta(2y - 1)$ under the accepting tie.

C TRAINING AND EVALUATION DETAILS

The joint arms use GRPO (Shao et al., 2024) to train Qwen3-4B-Instruct-2507 (Qwen Team, 2025) on HotpotQA (Yang et al., 2018) and MedQA (Jin et al., 2020). Rewards are centred within each question's group without standard-deviation scaling, as in Dr. GRPO (Liu et al., 2025), and sampling uses vLLM (Kwon et al., 2023).

Setting	Value
Samples per question	32
Questions per update	12
Updates; checkpoint interval	400; 100
Answer cap, training; evaluation	1,536; 4,096 tokens
Report cap, joint arms and trained reporters	32 tokens
Temperature	1.0
Learning rate; warm-up	10^{-6}; 12 updates
Clipping	0.2
KL term	None

Training records are checked every 50 updates through update 400. The cut share is the fraction of answers cut at the length cap. Over the preceding ten updates, a run fails if its cut or invalid-answer share exceeds 0.50, its invalid-report share exceeds 0.20 on a joint arm, or the share of groups with identical rewards exceeds 0.90. It also fails if the mean reward falls by more than three standard errors from its mean over the first ten updates. A HotpotQA run also fails if the mean answer length in that window is below 20 tokens, or the longest completion over the preceding twenty updates is below 64 tokens. The share of valid reports at exactly 0 or 1 is used only as a diagnostic. At update 200, an evaluation gate uses 200 questions disjoint from training and the held-out set. This gate rejects a run if its accuracy falls below the base model's by more than three standard errors of the paired difference, or its invalid-answer share exceeds 0.50. Both gates must pass before held-out scoring.

Readers are fitted on 2,000 questions disjoint from the held-out set. These questions may overlap with those used to train the answerer. The probe and classifiers use saved features of the frozen correctness-only checkpoint, while recalibration fits a logistic map from its reports to correctness. The verbalised report and recalibration baselines use a 1,536-token report cap.

The LoRA (Hu et al., 2022) reporters train on one stored answer per fitting question, generated by the frozen correctness-only checkpoint at the evaluation cap. Each group samples 32 reports on that answer, using the adapter only for the report. They use the same questions per update and update count as above, with a learning rate of 10^{-5}, making 2.4 passes over the store. The reward is the stored answer's correctness minus the confidence term λS_η of the chosen setting. The matched full-Brier reporter uses RLCR's reward on the same stored answers.

The advantage-variance share reported in Section 4.4 is the pooled share of a run: the covariance of the centred advantage with its correctness part, summed over groups, divided by the summed advantage variance, so that each group counts by what its update carried; the confidence term's share is one minus it. Each run's share uses all recorded completions through its last available update at the 24 September snapshot, with run-level shares averaged over seeds. The prefix-sharing implementation computes each question-and-answer prefix once per report micro-batch and preserves the objective and effective batch while changing the order of floating-point operations. In step-level checks, the report-token log-probabilities used in the loss agreed with the plain path to a maximum absolute error of 2.8×10^{-4} nats in float32 and within rounding noise in bf16. Seven runs on the original implementation were retained as matched controls, and the reported reporter results use the prefix-sharing runs.

Joint runs use four cards with a 24-hour wall limit. Short timing tests, or training smokes, project 12.9 hours and 51.7 GPU-hours per HotpotQA run, and 14.6 hours and 58.5 GPU-hours per MedQA run. The stored-answer reporters each use one GH200 card. In the limit setting, they use micro-batches of 8 on HotpotQA and 16 on MedQA, with a 12-hour wall limit.

Evaluation uses the fixed held-out set of 1,000 HotpotQA questions and MedQA's 1,273-question test part. The identity checks compare each run's model, task, reward setting, seed slot and resolved recipe with the design. They also trace table inputs and fitted readers to their checkpoints. MedQA's test file must match the published release's hashes and row count. Every arm starts with the same three random seed draws, and replacement runs retain their original slots for the paired comparisons.

Two-method comparisons pair questions within common seed slots and use a 10,000-resample bootstrap for the intervals. Seed means and per-seed figures retain the available cohort, with incomplete cohorts marked. The paper's tables include three seeds for the HotpotQA limit-setting middle arm. The earlier two-seed summary in the results file excludes seed 5506.

D Additional results

Table 5 gives the costs of each reader and reporter at every evaluation band on HotpotQA (Yang et al., 2018) and MedQA (Jin et al., 2020); Section 4.2's Table 3 shows its HotpotQA readers. For the composite rows of Table 2, each band column uses that band's own reader or reporter, and the whole-range columns use the middle one.

Table 6 lists individual-seed results for the joint limit-setting arms, RLCR and the equal-weight band readers.

		HotpotQA				MedQA			
Method	Fitted band	Low	Middle	High	$[0, 1]$	Low	Middle	High	$[0, 1]$
Answer probability	–	0.445	0.429	0.292	0.360	0.243	0.244	0.207	0.213
Verbalised report	–	0.443	0.426	0.377	0.397	0.243	0.241	0.223	0.224
Linear probe / BCE classifier	–	0.409	0.327	0.207	0.301	0.250	0.237	0.175	0.205
Brier classifier	–	0.410	0.327	0.206	0.301	0.248	0.238	0.179	0.206
Recalibration	–	0.432	0.375	0.236	0.327	0.245	0.237	0.190	0.208
Full-Brier reporter	–	0.428	0.346	0.224	0.315	0.243	0.232	0.174	0.201
Band reader, limit setting	Low	0.406	0.330	0.212	0.302	0.244	0.306	0.249	0.240
Band reader, limit setting	Middle	0.406	0.324	0.202	0.297	0.256	0.245	0.192	0.214
Band reader, limit setting	High	0.408	0.322	0.202	0.298	0.327	0.269	0.179	0.244
Band reader, equal-weight	Low	0.406	0.327	0.208	0.300	0.245	0.296	0.243	0.236
Band reader, equal-weight	Middle	0.406	0.324	0.202	0.297	0.255	0.244	0.191	0.213
Band reader, equal-weight	High	0.408	0.322	0.200	0.297	0.302	0.258	0.177	0.231
Band reporter, limit setting	Low	0.430	0.381	0.302	0.360	0.243	0.234	0.174	0.202
Band reporter, limit setting	Middle	0.425	0.348	0.224	0.315	0.243	0.233	0.172	0.201
Band reporter, limit setting	High	0.447	0.382	0.225	0.330	0.244	0.232	0.168	0.200
Band reporter, equal-weight	Low	0.420	0.347	0.222	0.313	0.244	0.235	0.178	0.204
Band reporter, equal-weight	Middle	0.425	0.349	0.224	0.315	0.243	0.233	0.174	0.201
Band reporter, equal-weight	High	0.449	0.411	0.240	0.343	0.244	0.232	0.170	0.201

Table 5: Mean decision costs for the 4B readers and reporters over three seeds, with each fitted band shown separately. Lower is better in every numeric column. The separate rows show which fitted reader or reporter supplies each composite cell.

For the MedQA comparisons in Section 4, Table 4 in Section 4.3 gives the paired cost differences. Section 4 reports mean comparisons of the band reader against RLCR by band, and of the limit-setting reporters against RLCR and the equal-weight reporters.

Table 7 presents the equal-weight joint arms from the ablation in Section 4.4, with RLCR as the reference.

Method and fitted band	Seed	Accuracy	Cost low	Cost middle	Cost high	Brier	AUROC
HotpotQA							
RLCR	3689	0.562	0.411	0.338	0.212	0.172	0.821
	6754	0.562	0.410	0.329	0.203	0.162	0.837
	7349	0.563	0.408	0.332	0.213	0.169	0.817
Limit setting, low	3689	0.542	0.415	0.376	0.247	0.195	0.805
	6754	0.558	0.409	0.344	0.249	0.199	0.782
	7349	0.545	0.424	0.356	0.221	0.177	0.834
Limit setting, middle	3689	0.539	0.421	0.336	0.216	0.159	0.839
	3952	0.542	0.409	0.328	0.207	0.145	0.858
	5506	0.499	0.436	0.344	0.216	0.138	0.875
Limit setting, high	1874	0.440	0.499	0.389	0.234	0.158	0.849
	3689	0.504	0.450	0.358	0.215	0.156	0.851
	6754	0.457	0.497	0.372	0.212	0.158	0.855
Band reader, equal-weight, low	3689	0.557	0.404	0.324	0.206	0.153	0.856
	6754	0.546	0.410	0.332	0.208	0.152	0.859
	7349	0.560	0.403	0.326	0.208	0.156	0.847
Band reader, equal-weight, middle	3689	0.557	0.404	0.321	0.200	0.148	0.861
	6754	0.546	0.409	0.328	0.204	0.148	0.860
	7349	0.560	0.404	0.322	0.200	0.151	0.856
Band reader, equal-weight, high	3689	0.557	0.407	0.322	0.199	0.150	0.860
	6754	0.546	0.411	0.324	0.202	0.146	0.863
	7349	0.560	0.404	0.319	0.199	0.149	0.861
MedQA							
RLCR	3689	0.775	0.226	0.220	0.178	0.161	0.703
	6754	0.756	0.243	0.235	0.187	0.168	0.736
	7349	0.782	0.217	0.216	0.184	0.163	0.752
Limit setting, low	3689	0.756	0.243	0.243	0.243	0.233	0.502
	6754	0.768	0.231	0.229	0.191	0.170	0.740
	7349	0.757	0.245	0.240	0.212	0.197	0.680
Limit setting, middle	3689	0.778	0.217	0.211	0.174	0.153	0.735
	6754	0.777	0.220	0.216	0.191	0.176	0.745
	7349	0.782	0.217	0.213	0.186	0.162	0.635
Limit setting, high	3689	0.761	0.234	0.221	0.163	0.146	0.787
	6754	0.760	0.226	0.212	0.164	0.138	0.793
	7349	0.717	0.250	0.216	0.168	0.123	0.851
Band reader, equal-weight, low	3689	0.754	0.248	0.299	0.240	0.230	0.701
	6754	0.774	0.229	0.288	0.244	0.230	0.740
	7349	0.742	0.258	0.300	0.245	0.227	0.743
Band reader, equal-weight, middle	3689	0.754	0.261	0.251	0.193	0.189	0.709
	6754	0.774	0.238	0.231	0.187	0.178	0.732
	7349	0.742	0.265	0.251	0.193	0.181	0.731
Band reader, equal-weight, high	3689	0.754	0.307	0.262	0.179	0.223	0.739
	6754	0.774	0.295	0.249	0.173	0.228	0.702
	7349	0.742	0.302	0.263	0.178	0.208	0.753

Table 6: Per-seed 4B results. Lower decision cost and Brier are better, as are higher accuracy and AUROC. Each row keeps the same run across all evaluation bands.

Task	Joint arm	Accuracy	Low	Middle	High	Brier	ECE	AUROC
HotpotQA	RLCR	0.562	0.410 (0.081)	0.333 (0.114)	0.209 (0.100)	0.168	0.067	0.825
	Equal-weight, low	0.556	0.434 (0.100)	0.383 (0.161)	0.243 (0.132)	0.216	0.076	0.680
	Equal-weight, middle	0.548	0.435 (0.096)	0.361 (0.135)	0.235 (0.122)	0.205	0.144	0.779
	Equal-weight, high	0.555	0.453 (0.119)	0.374 (0.151)	0.219 (0.107)	0.210	0.114	0.754
MedQA	RLCR	0.771	0.229 (0.057)	0.224 (0.109)	0.183 (0.126)	0.164	0.093	0.730
	Equal-weight, low	0.753	0.247 (0.062)	0.304 (0.180)	0.237 (0.176)	0.234	0.174	0.559
	Equal-weight, middle	0.767	0.233 (0.058)	0.231 (0.115)	0.205 (0.146)	0.192	0.116	0.603
	Equal-weight, high	0.756	0.245 (0.062)	0.244 (0.122)	0.193 (0.132)	0.176	0.083	0.712

Table 7: Mean results for the 4B equal-weight joint arms and RLCR over three seeds. The band columns give decision cost with regret in parentheses; lower cost, regret, Brier and ECE are better. A lower ECE need not accompany a lower band cost.