

Calibrated Is Not Informative: What a Fall in Expected Calibration Error Shows About LLM Confidence, Why Verbalised Scores Fail on Resolution as Well as Level, and the Recalibration Baseline the Ensembling Results Omit

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A common finding in work on language-model confidence is that a model asked to state its own confidence in one pass is badly overconfident, and that some added procedure - sampling several answers, a separate critique pass, a longer chain of thought - repairs it. The repair is usually scored by expected calibration error (ECE). A 2026 telecommunications study is a clear recent case: it reports that averaging several blind self-evaluations of a reasoning chain cuts ECE by up to 88 percent on three telecom benchmarks, and it reports no discrimination metric. In this analysis the author reviews and synthesises published results to show that such claims mix two failures that need different fixes. One is level: the scores sit too high on average. The other is resolution: the scores barely separate right answers from wrong ones. ECE mostly measures level. It can reach zero for a constant score that separates nothing, and a monotone map fitted on held-out labels can remove most level error without changing the ranking at all. In the published studies that report both ECE and a ranking metric, the two move independently. Better prompts cut ECE while, in the authors' summary, AUROC stayed close to chance. Verbalised scores cut ECE on one benchmark by three quarters while the area under the selective accuracy curve rose by about one point. Sampling several answers raised AUROC from 54.8 to 92.7 on arithmetic but by one point on a yes-no reasoning set, and a separate self-evaluation pass, the mechanism the telecom method repeats, was at chance on the one multiple-choice set where it was measured alone. Re-reading the telecom study's own tables, the author finds that its top confidence tier was about three points more accurate than the pooled base rate, and that its largest ECE reduction comes from a 150-item subset, where binning noise alone is close to the reported post-method value. The paper gives a decision procedure and five tests that would separate what added inference passes buy from what a cheap recalibration already buys.

Keywords: large language models; confidence estimation; calibration; expected calibration error; failure prediction; verbalized confidence; selective prediction; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A deployed language model is increasingly asked to do two things at once: answer, and say how far its answer can be trusted. The second job is what lets an operator route an answer to a human, to a retrieval step or to an automated action. In high-consequence settings it is the part of the system that carries the risk. The telecommunications case makes this concrete. Models are now proposed for reading 3GPP and O-RAN specifications, diagnosing network faults and generating configurations, and an operator who lets a model push a configuration without review is relying entirely on the model's confidence being meaningful (Saenko et al., 2026).

The published picture of verbalised confidence, a number or phrase the model emits about its own answer, is mixed in an instructive way. One line of work finds it badly overconfident. When asked for a percentage, models put most answers between 80 and 100 percent, often in multiples of five, with accuracy in each bin well below the stated level (Xiong et al., 2023). Vision-language models behave similarly (Groot and Valdenegro-Toro, 2024). Reasoning models often state

confidence above 85 percent, particularly when they are wrong (Mei et al., 2025). Another line finds the opposite. For several models fine-tuned from human feedback, verbalised probabilities are better calibrated than the models' own token probabilities, often by about half in relative ECE (Tian et al., 2023). Long chains of thought improve how well reasoning models express confidence, in 33 of 36 settings in one benchmark (Yoon et al., 2025). And a GPT-3 model can be trained to state well-calibrated confidence in words (Lin et al., 2022).

A third line takes the overconfidence as given and proposes a repair. Sampling several answers and measuring their agreement (Wang et al., 2022; Xiong et al., 2023; Lyu et al., 2024), asking the model to judge its own proposed answer as a separate step (Kadavath et al., 2022; Chen and Mueller, 2023), and extending the chain of thought before the confidence is stated (Podolak and Verma, 2025) are all reported to improve confidence. A 2026 study applies the second idea, repeated and averaged, to telecom question answering. The model answers with reasoning in a first pass. In a second pass it rates that reasoning several times, blind to the ground truth, on a 1-to-5 scale, and the ratings are averaged. The study reports ECE reductions of 17 to 88 percent across three Gemma-3 sizes and three telecom

benchmarks, and it describes the result as showing "exceptional discriminative power" (Saenko et al., 2026).

That phrase marks the problem this paper addresses. ECE does not measure discrimination. It measures how far stated confidence sits from observed accuracy within confidence bins, and a score that assigns every answer the same confidence equal to the overall accuracy has an ECE of zero while separating nothing. Tian et al. (2023) state the point in a footnote: in binary classification "a system that guesses randomly and outputs 50% confidence each time has perfect ECE". The image-classification literature has shown that the two properties can be pulled apart, and that most methods built to improve calibration do not help, or actively hurt, the task of flagging wrong predictions (F. Zhu et al., 2023; Jaeger et al., 2022). The survey literature on language-model confidence lists ECE, AUROC and accuracy-rejection areas as alternative ways to assess a score (Geng et al., 2024). In practice the literature mostly reports ECE. Sometimes it adds AUROC, and sometimes a selective-prediction AUC that is not AUROC. Its headline claims rarely say which property improved.

In this analysis the author reviews and synthesises published results to show that "verbalised confidence fails, and method X repairs it" contains two separate claims. The first concerns level: the scores are too high on average. The second concerns resolution: the scores carry little information about which answers are wrong. Level error is cheap to fix. A monotone map fitted on a modest labelled held-out set removes most of it and changes no ranking (Guo et al., 2017; Zadrozny and Elkan, 2002; Kumar et al., 2019). Resolution cannot be added by any function of the score alone; it needs new information, which is what sampling, a second pass or a longer chain of thought may supply. A method that costs several extra inference passes therefore earns its cost only through resolution, or through a level that holds across domains where a fitted map would not. ECE alone cannot show either.

The paper makes five contributions. First, it separates level, resolution and granularity as properties of a confidence score, and separates the two uses a score is put to: reading it as a probability, and thresholding it as a gate (Section 2). Second, it states several elementary facts about ECE that the language-model literature keeps running into, including a bound that turns out to bear on one of the telecom study's own numbers (Section 3). Third, it sets the evidence that verbalised confidence fails beside the evidence that it works, and shows that the two sides largely measure different properties (Sections 4 and 5). Where both properties are reported, they move independently (Section 6, Table 1, Figure 1). Fourth, it examines what added inference passes buy in discrimination, where it has been measured, and finds that the large gains come from resampling answers, while a repeated judgement of one fixed answer, which is what the telecom method uses, has weak measured support on multiple-choice questions (Section 7). It then re-reads the telecom result against its own tables

(Section 8). Fifth, it names the omitted baseline, a recalibrated single-pass score. It argues that a confidence used as an auto-accept gate should be judged by error at a given coverage, with a guarantee where one is available, rather than by ECE (Section 9). It closes with a decision procedure (Algorithm 1) and five tests (Section 12).

The argument is not that the telecom study, or any other cited study, is wrong about what it measured. Its ECE values are what they are. The argument is that ECE values cannot support the conclusion most often drawn from them, which is that a method makes a model better at knowing when it is wrong.

2. Scope and terms

The paper concerns confidence scores attached to individual answers of a language model in question answering, mostly with a short or multiple-choice answer whose correctness can be checked. It does not concern calibration of the full token distribution, long-form factuality at the level of claims, or confidence conveyed in hedging language that is never turned into a number. These are related, but the arguments here are made for the single-answer case. Where a cited result comes from image classification, that is stated.

Score. A number $s(x)$ attached to the model's answer on input x , meant to indicate how likely the answer is to be correct. Verbalised scores are produced as output tokens. Consistency scores are computed from the agreement among several sampled answers. Self-evaluation scores are produced when the model, in a separate step, judges an answer it has already given. Probe scores are read from hidden states by a trained classifier.

Level (calibration-in-the-large). The gap between the mean score and the overall accuracy on a dataset. A score with mean 0.92 on a dataset where the model is right 70 percent of the time has a level error of 0.22 whatever else is true of it.

Reliability (calibration in the narrow sense). Whether, among answers given score c , the fraction correct is c , for every c . Level error is the average of reliability error; a score can have zero level error and poor reliability if it is too high in one region and too low in another. Calibration in this sense is the property the classical forecast-verification literature defines, together with refinement (DeGroot and Fienberg, 1983).

Resolution (discrimination, refinement). How much the probability of being correct varies across the values the score takes. A score with high resolution gives right answers systematically higher values than wrong ones. Murphy's partition of the Brier score (Brier, 1950) makes the separation exact: the mean squared error of a probability forecast splits into a reliability term, which is a penalty, a resolution term, which is a credit, and an uncertainty term set by the base rate alone (Murphy, 1973). The modern statement of the same idea is that a probabilistic forecast should be as sharp as possible subject to being calibrated (Gneiting et al., 2007).

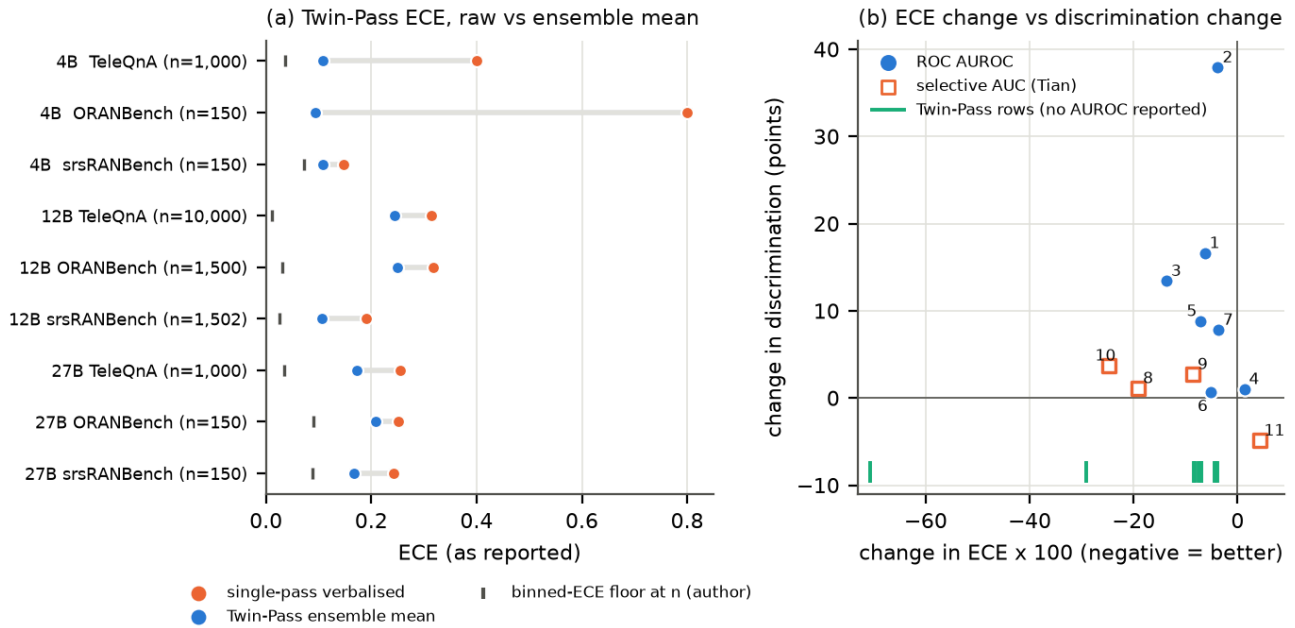


Figure 1. (a) ECE of single-pass verbalised confidence (orange) and of the Twin-Pass ensemble mean (blue) for each Gemma-3 model and telecom sub-benchmark, from Saenko et al. (2026), Tables I to III, with the number of items. Grey ticks mark the author’s approximate expected binned ECE of a perfectly calibrated score at that sample size (ten equal-mass bins, normal approximation, bin accuracy set to the reported accuracy); the source does not state its binning, so the ticks are indicative. (b) Change in ECE (multiplied by 100) against change in discrimination for interventions that report both, numbered as follows: 1 Xiong et al. (2023) GPT-3.5 five-dataset average, one CoT answer to five samples; 2 the same on GSM8K; 3 the same on professional law; 4 the same on StrategyQA; 5 Top-K against CoT prompting, GPT-3.5 average; 6 Pair-Rank against consistency aggregation, GPT-4 average; 7 Yoon et al. (2025) GLM-Z1 against GLM-4 on NonambigQA; 8 to 10 Tian et al. (2023) gpt-3.5 label probability to verbalised top-4 on SciQ, TriviaQA and TruthfulQA; 11 Tian et al. (2023) one-stage top-1 to two-stage CoT on TriviaQA. Filled circles are ROC AUROC; hollow squares are selective accuracy-coverage AUC, a different quantity. Green ticks at the bottom mark the nine Twin-Pass ECE changes, for which no discrimination metric is reported.

Granularity. The number of distinct values a score actually takes. A verbalised score on a 1-to-5 scale has at most five, and if the model uses only 4 and 5 it has two. Granularity caps resolution, because items given the same value cannot be told apart by the score.

ECE. Expected calibration error, the weighted mean over confidence bins of the absolute difference between accuracy and mean confidence in the bin (Naeini et al., 2015; Guo et al., 2017). It estimates reliability error, and through it level error. It is not a proper scoring rule and has no resolution term.

AUROC. The area under the receiver operating characteristic curve of the score as a detector of correct answers: the probability that a randomly chosen correct answer scores higher than a randomly chosen wrong one, with ties counted as one half (Hanley and McNeil, 1982; Fawcett, 2006). It is unaffected by any strictly increasing transformation of the score and ignores level entirely. It is the standard metric of failure prediction (Xiong et al., 2023; Jaeger et al., 2022).

Selective AUC. The area under the curve of accuracy against coverage when the lowest-scoring answers are withheld, the summary used in selective classification (Geifman and El-Yaniv, 2017). Unlike AUROC, it depends on the base accuracy as well as on ranking: a random score gives a selective AUC equal to the accuracy, not 0.5. Tian et al. (2023) and Shrivastava et

al. (2023) report this quantity under the name AUC. Reading it as AUROC overstates how much the score discriminates, which matters below.

Two uses. A score can be read as a probability, by a person or a downstream decision rule that multiplies it by a cost; then reliability matters. It can be thresholded as a gate, to auto-accept, retrieve or escalate; then what matters is the error rate among the answers that pass the gate and the fraction that pass. Jaeger et al. (2022) argue that the two requirements are distinct and that "either of the tasks can be solved without solving the other". A threshold can be set on a validation set using ranking alone, and calibration is then not needed at any stage. Chow’s classical analysis of the reject option already ties the optimal threshold to a ratio of costs, not to a calibration statistic (Chow, 1970).

3. Five facts about ECE

The following are derivations or published measurement results, not new findings. They are collected because the confidence literature for language models repeatedly runs into each of them.

Fact 1: a constant score at the base rate has zero ECE and no discrimination. If every answer receives score A , where A is the overall accuracy, all answers fall into one bin whose accuracy is A and whose mean confidence is A , so ECE is zero. Every pair of answers is tied, so

AUROC is exactly 0.5. Xiong et al. (2023) report an observed case close to this. With chain-of-thought prompting on GSM8K, GPT-4 reached 93.6 percent accuracy and an ECE of 0.064 by assigning 100 percent confidence to every sample, so that the scores could not distinguish correct from incorrect answers.

Fact 2: ECE is bounded below by the level gap, and above by the lowest confidence used. Because ECE is a weighted mean of per-bin absolute gaps, the triangle inequality gives $ECE \geq |\text{mean score} - \text{accuracy}|$. In the other direction, suppose every score lies in $[L, 1]$ and the overall accuracy is A . For a bin with accuracy a and mean confidence c in $[L, 1]$, the gap $|a - c|$ is at most $\max(1 - a, a - L)$. That function of a is convex, so averaging over bins whose accuracies have mean A gives $ECE \leq (1 - A) + A(1 - L) = 1 - A L$. For example, if every score is at least 0.75 and accuracy is 0.693, ECE cannot exceed 0.48. Conversely, an ECE of 0.80 at accuracy 0.693 requires some scores at or below 0.29. Section 8 uses this bound.

Fact 3: a monotone recalibration map of one score, fitted on held-out labels, fixes level and leaves ranking unchanged. Post-hoc probability modification of this kind was among the methods that improved the poor calibration of earlier generative question-answering models (Jiang et al., 2021). Temperature scaling, Platt scaling, isotonic regression and histogram binning replace each score by an estimate of the accuracy at that score, fitted on labelled held-out data (Guo et al., 2017; Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005). Applied to one scalar confidence, such as a 1-to-5 rating, Platt scaling is a strictly increasing map and leaves AUROC exactly unchanged, and isotonic regression is monotone, so it can merge neighbouring values into ties but cannot reverse their order. Temperature scaling of multiclass logits is a different case. It keeps the argmax, and Guo et al. (2017) note that it therefore does not change the predicted class or the accuracy. But the rescaled top probability also depends on the other logits, so temperature scaling can reorder confidences across examples slightly. Histogram binning replaces each distinct score value by the held-out accuracy at that value. No function of the score can separate two answers that received the same score. The best AUROC any such function can reach is obtained by ordering the score values by their held-out accuracy, which is what binning does. Recalibration can therefore raise AUROC only up to the ceiling the score's own distinct values allow, and only where the original values were out of order. For a score with B distinct values, histogram binning needs on the order of B/ϵ^2 labelled samples to reach calibration error ϵ (Kumar et al., 2019). For a five-level verbalised score this is a small number.

Fact 4: binned ECE is a biased, noisy estimator, and the bias is large at small sample sizes. The usual plug-in estimator with equal-width bins is biased, and the bias depends on the number of samples and bins. Equal-mass bins and debiased estimators do better (Roelofs et al.,

2022). Scaling-based recalibrators were found to be less calibrated than their plug-in ECE suggested (Kumar et al., 2019). Rankings of recalibration methods change with the choice of calibration measure (Nixon et al., 2019), and ECE is not even a continuous function of the predictor (Blasiok et al., 2023). For a rough sense of scale, a perfectly calibrated score evaluated on n items in ten equal-mass bins has an expected per-bin gap of about $\sqrt{2p(1-p)/(\pi n/10)}$ under a normal approximation, where p is the accuracy in the bin. At $p = 0.7$ that is about 0.09 for $n = 150$, 0.04 for $n = 1,000$ and 0.01 for $n = 10,000$. These are the author's arithmetic under the stated assumptions, not values any cited study reports. They show that on 150 items an ECE near 0.09 is what sampling alone would produce from a perfect score.

Fact 5: calibration estimated on one distribution need not hold on another. In a large benchmark of image and other classifiers under dataset shift, traditional post-hoc calibration fitted in distribution fell short once the inputs shifted (Ovadia et al., 2019). For language models, a P(K) classifier trained only on TriviaQA reached an AUROC of 0.606 and a Brier score of 0.431 on LAMBADA. Training on a mixture of tasks raised these to 0.853 and 0.108 (Kadavath et al., 2022). For pre-trained transformers, temperature scaling further reduced calibration error in domain, while out of domain a different remedy, label smoothing, was what helped (Desai and Durrett, 2020). Thermometer was built because a separate fitted temperature for each new task is impractical for a general-purpose model. It learns, from data on many tasks, an auxiliary model that calibrates the language model on tasks it has not seen (Shen et al., 2024). Fact 5 is the strongest argument for a confidence method that needs no labels, and Section 9 returns to it.

4. The case that verbalised confidence fails

The negative evidence is broad and largely consistent on level. What it shows about resolution is weaker, and sometimes the opposite.

Xiong et al. (2023) give the most systematic black-box account. They evaluated five models on eight datasets covering arithmetic, commonsense, symbolic, ethical and professional-knowledge questions. Their vanilla verbalised confidence produced average ECE of 0.520 for GPT-3, 0.377 for GPT-3.5 and 0.180 for GPT-4, with average AUROC of 51.3, 55.1 and 62.7 percent. The confidence values clustered between 80 and 100 percent in multiples of five, which the authors relate to the way humans talk about confidence in the training corpus. Their Table 2 thus shows both failures for GPT-3 and GPT-3.5: high ECE and AUROC near chance. For GPT-4, ECE fell by more than half relative to GPT-3.5 while average AUROC rose by 7.6 points, and in their words AUROC remained close to the random threshold. Most of the gain from scale was in level.

Several studies find the same overconfidence with different instruments. Across four language models and two vision-language models, prompted verbalised

uncertainty was mostly overconfident, and a signed calibration measure introduced for the purpose showed the direction (Groot and Valdenegro-Toro, 2024). Publicly deployed models seldom expressed uncertainty unless asked. When explicitly asked they leaned overconfident, and on average 47 percent of the answers they marked as confident were wrong (Zhou et al., 2024). A metric of faithful response uncertainty compares a model's internal consistency with the decisiveness of its wording, and on it aligned models conveyed their uncertainty poorly (Yona et al., 2024). Reasoning models, studied across a range of benchmarks, usually gave confidence above 85 percent, and more so on answers that turned out wrong (Mei et al., 2025). In a benchmark of verbalised confidence across many models and prompt designs, models of 70 billion parameters and above had ECE around 0.1. The authors attribute most of the improvement with size to higher accuracy rather than to reduced overconfidence, and for the smallest model they tested the scores were nearly independent of accuracy (Yang et al., 2024).

Two accounts of why the level is high have support. One is imitation. Certainty and uncertainty markers placed in prompts shift a model's accuracy by large amounts. In pretraining text, markers of uncertainty tend to occur in answers on question-answering sites and markers of certainty in questions, which suggests that models copy patterns of usage rather than report an epistemic state (Zhou et al., 2023). The other is post-training. The GPT-4 technical report shows that the pre-trained model's answer probabilities on an MMLU subset were close to calibrated and that post-training reduced calibration markedly (OpenAI et al., 2023). Kadavath et al. (2022) found RLHF policies miscalibrated in the same way, and found that a single temperature of 2.5 largely restored calibration on several evaluations. Reward models used in RLHF prefer responses that state high confidence regardless of quality, which pushes verbalised confidence upward (Leng et al., 2024). Ordinary reinforcement learning with a binary correctness reward degrades calibration, because it never penalises confident guessing (Damani et al., 2025). Systematic studies of alignment and calibration trace the effect through pre-training and alignment stages (C. Zhu et al., 2023).

Both accounts explain a level shift. Neither predicts that the ordering of scores carries no information. The RLHF case shows this most clearly. Kadavath et al. (2022) restored calibration with a single temperature applied to the policy's distributions over answer options. That adjustment keeps each question's top option and adds no information the policy's probabilities lacked; whatever ranking information they held was there before it, and what the adjustment mainly moved was the level. Studies whose case against verbalised confidence rests on ECE alone have, by Fact 3, documented a level problem. They have not by that alone documented a resolution problem.

Resolution failures are documented where AUROC is reported. Xiong et al. (2023) report AUROC near chance

for vanilla verbalised confidence from GPT-3, Vicuna and GPT-3.5. Kapoor et al. (2024) argue that prompting alone does not give useful uncertainty for open-source models on an open-ended variant of MMLU. Their figures report both ECE and AUROC, and a small fine-tuning set of about a thousand graded examples improves both. Both papers show the resolution failure directly, not by inference from ECE.

5. The case that verbalised confidence works

The positive evidence is also real, and is also mostly evidence about level.

Tian et al. (2023) evaluated ChatGPT, GPT-4, two Claude versions and Llama-2-70B-Chat on TriviaQA, SciQ and TruthfulQA. For gpt-3.5-turbo on SciQ, the model's own label probability had ECE 0.256. A verbalised probability elicited with four candidate answers had ECE 0.065, a relative reduction of 75 percent. The selective accuracy-coverage AUC changed from 0.752 to 0.763. On TriviaQA the corresponding ECE values were 0.140 and 0.054, and the AUC rose from 0.869 to 0.896. On TruthfulQA ECE fell from 0.451 to 0.203 and AUC rose from 0.418 to 0.455. In every case ECE fell by more than half while the ranking-sensitive quantity moved by one to four points. The study is careful on the point this paper presses. It reports ECE after temperature scaling alongside raw ECE, fitting the temperature on a fifth of the data. That shrinks the gap: label probability on SciQ falls from 0.256 to 0.180 after scaling. The authors also say in the text that ECE does not capture discriminative power. Tian et al. (2023) is thus the rare study that runs the recalibration baseline this paper argues for. Under that baseline the verbalised advantage on level shrinks but does not vanish.

Lin et al. (2022) fine-tuned GPT-3 to state its confidence in words, and found the stated levels well calibrated and moderately robust under distribution shift. Mielke et al. (2022) found chat models poorly calibrated in their expressions of doubt. They also found that the probability of a correct answer could be predicted from other features, and used such predictions to train a controllable model whose linguistic confidence matched. Band et al. (2024) extended the idea to long-form text with a reinforcement-learning objective defined on what a reader can predict. Damani et al. (2025) trained reasoning models with a reward that adds a Brier score to binary correctness. They proved that a bounded proper scoring rule makes accurate and calibrated predictions optimal, and reported better calibration with no loss of accuracy, in and out of domain. Because the Brier score has a resolution term (Murphy, 1973), this approach rewards resolution as well as level, unlike training against ECE.

Two results bear directly on resolution. Yoon et al. (2025) compared six reasoning models with non-reasoning counterparts of the same backbone on six datasets. The reasoning models were strictly better calibrated in 33 of 36 settings, and the advantage was traced to slow-thinking behaviours that let confidence

adjust as the chain unfolds. Their metrics include AUROC. In one example, a reasoning model with 0.07 lower accuracy than its non-reasoning counterpart had 0.037 lower ECE and 0.078 higher AUROC, which the authors read as evidence that better confidence is not just a by-product of accuracy. Podolak and Verma (2025) forced DeepSeek R1-32B to reason before stating confidence. A budget of 100 to 500 reasoning tokens raised accuracy from 41 to 63 percent and verbalised-confidence ROC-AUC from 0.56 to 0.80, reaching 0.88 with longer reasoning. A separate reader model given only the chain of thought matched the model's ROC-AUC (0.841 against 0.851), with a moderate rank correlation of 0.57 between the two scores. The authors infer that the stated number summarises the alternatives the chain has surfaced. These are resolution gains measured by a ranking metric, and they support the idea behind the telecom second pass, that a model reading a reasoning trace can extract information about its correctness.

The results also contain warnings. Yoon et al. (2025) note cases where non-reasoning models had higher AUROC but worse Brier scores. They attribute these to implicit cues in multiple-choice options inflating discrimination without calibration, so even within one study the two metrics disagree on which model is better. Mei et al. (2025) found that reasoning models became more overconfident on answers they got wrong as reasoning effort increased. Calibration improved with depth where deeper reasoning raised accuracy, and worsened as accuracy saturated. Asking the model to reason about its own chain made two models less overconfident and a third more so. Those results are stated in calibration-error terms, and the paper reports no ranking metric for them. So "reasoning helps confidence" has measured support for resolution in two studies, measured support for level in several, and a measured counterexample for level in one.

The disagreement between the two literatures is therefore smaller than it looks. Verbalised confidence from instruction-tuned models has a large level error that varies by model and prompt. A modest share of that error survives fitted recalibration. Its resolution varies widely: near chance for older or smaller models and for knowledge-heavy questions, and substantial for strong models when the score follows explicit reasoning. "Verbalised confidence fails" is mostly true of level. "Verbalised confidence works" is mostly true of level after better prompting or training, and in places of resolution.

6. Where both are measured, they dissociate

The core evidence for this paper's thesis is the set of studies that report ECE and a ranking metric for the same intervention. Table 1 collects them, and Figure 1(b) plots the changes. Five patterns stand out.

First, ECE can fall without any gain in ranking. Xiong et al. (2023) found that human-inspired prompting reduced ECE for GPT-3.5 and GPT-4 while failure prediction stayed close to chance. They concluded that "a reduction

in overconfidence is due to the diminished gap between average confidence and accuracy, not necessarily indicating a substantial increase in the model's ability to judge the correctness of its responses". In Tian et al. (2023), the switch from label probability to verbalised confidence cut ECE by 55 to 75 percent on three datasets for gpt-3.5-turbo while the selective AUC moved by one to four points.

Second, ranking can rise without a fall in ECE. In Xiong et al. (2023), Table 3, moving GPT-3.5 from a single chain-of-thought response to five self-randomised samples aggregated by consistency left StrategyQA ECE slightly worse (22.0 to 23.3, on a scale of 100). AUROC rose by one point there and by 9 to 38 points on the other four datasets.

Third, the method that wins on one metric can lose on the other. Comparing aggregation strategies for GPT-4, Xiong et al. (2023) found Pair-Rank best on average ECE (6.9 against 12.0 for plain consistency) while average AUROC was 67.6 against 66.9. A third aggregator, Avg-Conf, had the highest AUROC on five of six datasets and the same average AUROC as plain consistency, yet the worst average ECE of the three (14.8). Their summary is that one aggregator suits calibration and another failure prediction. Kadavath et al. (2022) report the same kind of trade: more examples in a few-shot self-evaluation prompt mainly improved calibration rather than AUROC, and in some cases the best-discriminating set-up was not the best calibrated.

Fourth, methods designed for calibration can damage ranking. In image classification, mixup, label smoothing and focal loss are known to lower calibration error, yet F. Zhu et al. (2023) found that "most confidence calibration methods are useless or harmful for failure prediction". On CIFAR-10 with ResNet110, baseline AUROC of 90.76 fell to 86.93 with mixup and to 81.89 with label smoothing. Post-hoc temperature scaling did not help either: fitted on validation data it slightly hurt failure prediction, and even fitted on the test set it gave a negligible gain (F. Zhu et al., 2023). A large benchmark of confidence-scoring functions under several shift types found the plain softmax response the best overall at detecting failures. Its authors argue that calibration should be treated as a separate task from ranking (Jaeger et al., 2022). A follow-up showed that the choice of multi-threshold metric changes method rankings on five of six datasets (Traub et al., 2024).

Fifth, ECE gains can track accuracy rather than confidence. When an intervention raises accuracy toward the level at which an overconfident model already reports itself, ECE falls even if the confidence scores do not change. Xiong et al. (2023) say this of chain-of-thought prompting. Yang et al. (2024) say it of model scale. Mei et al. (2025) find calibration improving with reasoning depth where depth raised accuracy. Any intervention that changes the answers as well as the scores confounds the two. The telecom second pass does not change the answers, which makes its ECE comparison cleaner on this point than most (Section 8).

The empirical reading of these patterns matches the theory: ECE and ranking metrics are close to independent descriptions of a score. A broader evaluation framework adds a third dimension, robustness of the score to rephrasing of prompts and answers, and finds it largely independent of the established metrics as well (Xia et al., 2026). Confidence of language models varies across semantically equivalent inputs (Mahaut et al., 2024). Large benchmarks of uncertainty methods for language models report several metrics side by side because no single one captures quality (Vashurin et al., 2024; Fadeeva et al., 2023). An ECE improvement, reported alone, is evidence about one of at least three properties.

7. What added inference passes buy

The repairs proposed for verbalised confidence all spend extra inference. They can be grouped by the information they add.

Sampling agreement. Several answers are drawn and their agreement is measured, by majority vote, entropy over answer clusters, or semantic equivalence (Wang et al., 2022; Kuhn et al., 2023; Manakul et al., 2023; Lyu et al., 2024). Agreement adds information about the model's answer distribution that a single sample does not reveal. Semantic entropy detects confabulations across datasets without task-specific tuning (Farquhar et al., 2024), and probes can approximate it from a single generation's hidden states (Kossen et al., 2024). The measured resolution gains are large where answers are open-ended and can differ. Xiong et al. (2023) report GSM8K AUROC rising from 54.8 to 92.7 for GPT-3.5 with five samples. On professional law the same change took AUROC from 52.2 to 65.6, and on StrategyQA from 59.8 to 60.8. Lyu et al. (2024) drew 40 samples at temperature 0.4 and scored results mainly with the Brier score. For GPT-4, averaged over domains and prompts, the three consistency measures gave 0.114 to 0.119, against 0.154 and 0.181 for two verbalised baselines elicited after the reasoning chain. Their "post-hoc" baselines are extraction methods that need no labels. None is a recalibration fitted to labelled data, so "outperforms post-hoc approaches" in that paper does not mean it outperforms temperature scaling or binning.

Self-evaluation in a separate pass. The model is shown its own answer, sometimes with reasoning, and asked whether it is correct. This is the mechanism of the telecom second pass. It has a long history. Kadavath et al. (2022) asked models for the probability that a proposed answer is true and found self-evaluation workable and well calibrated few-shot for large models. It improved further when the model was first shown several of its own samples to compare. They also noted that models tend to find their own samples more plausible. Weng et al. (2023) used backward verification of a derived answer to pick among candidates. Chen and Mueller (2023) combined sampling agreement with a self-reflection certainty, asking the model in follow-up questions whether its earlier answer was correct. Their

Table 1 lets the two parts be compared. For GPT-3.5-turbo, self-reflection alone gave AUROC 0.831 on GSM8K, 0.839 on SVAMP, 0.655 on TriviaQA and 0.506 on CommonsenseQA. The combined score gave 0.951, 0.927, 0.817 and 0.769. On the one multiple-choice dataset, separate-pass self-reflection was at chance, so the combined score's resolution must have come mainly from its sampling-based component, which the paper does not report separately.

The negative evidence on self-evaluation is extensive. Intrinsic self-correction without external feedback did not improve, and sometimes degraded, reasoning accuracy (Huang et al., 2024). Models struggled to locate logical mistakes even in unambiguous cases, though they corrected well once told where the mistake was (Tyen et al., 2024). On planning and constraint tasks, GPT-4's self-critique led to collapse while a sound external verifier gave large gains (Stechly et al., 2024). Model evaluators recognise their own outputs and score them higher (Panickssery et al., 2024). Self-preference appears in judge settings generally (Zheng et al., 2023; Wataoka et al., 2024). The generative AI paradox hypothesis, which reports weaker understanding than generation in several controlled comparisons, also cuts against the assumption that judging is easier than producing (West et al., 2023). The telecom study grounds its design in the opposite assumption, that models are better at critiquing a reasoning chain than at producing a correct answer, and cites Weng et al. (2023) and Podolak and Verma (2025) for it (Saenko et al., 2026). Both assumptions have published support. Which holds for a given model and domain is an empirical question that only a ranking metric can answer.

Cross-model evaluation. Several results find that a second model can judge the first. Kapoor et al. (2024) report that models fine-tuned as uncertainty estimators work on other models' answers as well as their own. Shrivastava et al. (2023) found that a weaker surrogate model's probabilities gave higher selective AUC than GPT-4's own linguistic confidence on 9 of 12 datasets. Mixing the two raised GPT-4's average from 80.5 to 83.4 percent, against a random baseline about seven points lower, which shows the metric is selective AUC and not AUROC. On AUROC, which the same study also reports, GPT-4 rose from 64.6 to 69.9. The telecom study reports that a 4B model critiquing 12B answers reached a top-tier accuracy near 73 percent, a tier it could not reach on its own answers (Saenko et al., 2026). That is consistent with the cross-model results and with the self-preference results, and it is the most interesting observation in that study.

Longer reasoning before the score. Covered in Section 5: measured ranking gains for one reasoning model (Podolak and Verma, 2025) and for reasoning models in one six-model comparison (Yoon et al., 2025), measured calibration losses at higher effort where accuracy had saturated, though calibration improved on average (Mei et al., 2025).

Training. Fine-tuning on graded examples (Kapoor et al., 2024), auxiliary confidence predictors trained on text alone (Ulmer et al., 2024), training with self-reflective rationales (Xu et al., 2024), listener-aware fine-tuning (Stengel-Eskin et al., 2024) and calibration-aware reinforcement learning (Damani et al., 2025; Leng et al., 2024) use labels to learn a score. Supervised hidden-state probes gave the most reliable factual confidence in one systematic comparison, at the cost of needing weights and training data (Mahaut et al., 2024). The alignment between a model's token probabilities and its stated confidence is itself only moderate, with GPT-4's average Spearman correlation at 0.42 in one study (Kumar et al., 2024), so the verbal channel and the probability channel are not interchangeable.

Two conclusions follow. Added passes can add resolution, and where it has been measured they often do, by amounts that vary widely by dataset: from one AUROC point on StrategyQA to 38 on GSM8K for the same sampling change, with the two MMLU multiple-choice subsets in between at 13 and 23 points (Xiong et al., 2023). Those gains come from sampling new answers. The telecom method does not resample the answer; it repeats a judgement of one fixed reasoning chain. For that mechanism in isolation, the only located decomposition on a multiple-choice set put it at chance, while on arithmetic it did well (Chen and Mueller, 2023). The ensembling repair therefore cannot be assumed to add resolution in the telecom setting. Only a ranking metric computed in that setting could show it.

8. Re-reading the telecom result

Saenko et al. (2026) evaluated Gemma-3 at 4B, 12B and 27B parameters (Gemma Team et al., 2025) on three sub-benchmarks from the GSMA Open Telco suite. TeleQnA draws its questions from 3GPP standards and research material and has 10,000 items in full (Maatouk et al., 2023). ORANBench derives from O-RAN Alliance specifications; its parent benchmark has 13,952 multiple-choice questions generated from 116 documents (Gajjar and Shah, 2024). srsRANBench targets an open-source RAN codebase. The 4B and 27B models were run on a lite set of 1,300 items (1,000, 150 and 150). The 12B model was run on the full set of 13,002 (10,000, 1,500 and 1,502). The baseline asks for an answer, reasoning and a 1-to-5 confidence in one call. The method keeps that answer and reasoning and asks the same model, in N separate calls, to rate the reasoning on the same 1-to-5 scale without the ground truth; the N ratings are averaged (ensemble mean) or their median is taken. Because the answer is unchanged, accuracy is identical for baseline and method, and every difference in ECE is a difference in the score.

The study reports, per model and sub-benchmark, the accuracy, the raw ECE, and the ECE of the ensemble mean and median (its Tables I to III; Table 1 and Figure 1(a) here). The relative ECE reductions with the mean run from 17.3 to 88.4 percent. It reports reliability diagrams pooled over sub-benchmarks, a figure showing

single-pass confidence for the 4B model clustered at 4 and 5 and "largely indistinguishable between correct and incorrect answers", and a "Golden Threshold Map" assigning ensemble-mean tiers to operational actions. That map gives accuracies of 0.73 and 0.78 for the 12B and 27B models in the top tier (ensemble mean 4.5 to 5.0, labelled auto-accept), 0.65 and 0.71 in the 4.0 to 4.5 tier, 0.59 and 0.66 in the 3.0 to 4.0 tier, and at most 0.44 and 0.60 below 3.0. The 4B model never reached the top tier. Mean aggregation beat median in 55 percent of conditions.

Five observations follow from the published numbers.

The paper reports no ranking metric. Its claim of "exceptional discriminative power" rests on reliability diagrams and on ECE, neither of which measures discrimination (Section 2). The single-pass figure for the 4B model shows the baseline's resolution failure. Nothing in the paper measures the method's resolution, so the improvement shown is an improvement in calibration error.

The top tier is only a little more accurate than the base rate. Assuming the threshold map pools the three sub-benchmarks, as the reliability diagrams do, the author computes pooled accuracy from Tables I to III as 0.698 for the 12B model (9,073 correct of 13,002) and 0.745 for the 27B model (968 of 1,300). The auto-accept tier's accuracies of 0.73 and 0.78 are therefore 3.2 and 3.5 points above what auto-accepting every answer would give. Lower tiers are less accurate, so the score does carry some ranking information; a bottom tier at or below 0.44 for the 12B model is a real separation. But an auto-accept rule whose accepted answers are wrong 22 to 27 percent of the time is a modest gate. The study does not report what share of answers falls in each tier. Without those shares, neither AUROC nor the error rate at a given coverage can be reconstructed from the paper.

The largest single reduction comes from a 150-item subset. The 88.4 percent reduction is for the 4B model on the 150 ORANBench items. The other three 150-item reductions, 17.3, 26.5 and 31.0 percent, fall within or below the range of the full-set reductions. On the full sets, where the 12B model was run on 10,000, 1,500 and 1,502 items, the reductions are 22.2, 21.6 and 43.9 percent, and the ensemble mean's ECE remains 0.244 on TeleQnA and 0.249 on ORANBench. Model size and set size are confounded in this design, so this does not show that the method works less well at 12B. It does show that the ensemble mean, on two of the three largest samples in the study, left the score with an ECE near a quarter. By the arithmetic of Fact 4, that is far above the noise floor at those sizes (about 0.01 to 0.03). On the 150-item sets, the post-method ECE values of 0.093 for the 4B model on ORANBench and 0.108 on srsRANBench are of the same order as the 0.07 to 0.10 that binning noise alone would produce from a perfectly calibrated score, under ten equal-mass bins (Figure 1(a)). The study states neither its binning scheme nor its number of bins, so this comparison is indicative only. On this subset the post-method ECE of 0.093 cannot be distinguished from zero,

so the information published does not show how well calibrated the ensemble mean is there, only that ECE fell from a very high raw value. Whether that raw value of 0.800 is itself sound is the next question.

One raw ECE cannot be reconciled with the paper's description of the scores. The 4B model on ORANBench has accuracy 0.693 and raw ECE 0.800. By Fact 2, if every single-pass score had been 4 or 5, mapped onto $[0, 1]$ as either $(s - 1)/4$ or $s/5$, ECE could not have exceeded 0.48 or 0.45. An ECE of 0.800 at that accuracy requires some confidences at or below 0.29 on whatever scale was used, which on a 1-to-5 scale means ratings of 1 or 2 under the first mapping. That is compatible with the data only if the ORANBench subset differs sharply from the pooled picture the paper describes, if unparseable outputs were scored at the bottom of the scale, or if a mapping onto $[0, 1]$ the paper does not state was used. The paper does not give its mapping. This is the single largest raw value in the study, and it produces the 88.4 percent headline. It may well be correct, but it cannot be reconstructed from what is published.

Key settings are not reported. The number of evaluation passes N , the sampling temperature of the evaluation calls, the mapping from the 1-to-5 scale to $[0, 1]$, the number and kind of ECE bins, and whether the threshold map was fitted and evaluated on the same items are all absent. The illustrative ensemble values in the paper's text (3.33, 3.67 and 4.25) and in its appendix (1.66) are consistent with more than one N .

None of this shows that the twin-pass method fails to improve confidence in the telecom setting. The method may well add resolution: the self-evaluation and cross-model results in Section 7 give reason to expect some, and the tier accuracies show some ranking. The point is that the paper's evidence establishes a reduction in calibration error, partly on samples too small to estimate it precisely. It does not establish the property its conclusion relies on, which is that a high ensemble score marks an answer safe to act on.

9. The omitted baseline, and the right target for a gate

A confidence method that costs $N + 1$ inference calls per answer should be compared with the cheapest alternative that achieves the same thing. For level, that alternative is a recalibrated single-pass score.

The baseline. Take the single-pass 1-to-5 rating and replace each of its values with the accuracy observed at that value on a labelled held-out split. This is histogram binning with at most five bins (Zadrozny and Elkan, 2002). Its sample cost is proportional to the number of distinct values (Kumar et al., 2019), and with about a hundred labelled items per value, each per-value accuracy has a standard error of about five points. In distribution, its expected ECE is close to the binning noise floor of Fact 4. It costs no extra inference. By Fact 3, its AUROC is at most what ordering the five values by

accuracy gives. If the model uses only 4 and 5, that is the resolution of a two-valued score. The telecom study already had the labels needed to fit such a map, since it computed ECE and built its threshold map from them.

The comparison this baseline makes possible has three outcomes. If the ensemble mean and the recalibrated single pass have similar ECE in distribution, the ensemble's ECE advantage over the raw single pass is a level effect a five-parameter map buys for free. If the ensemble has higher AUROC, or lower error at a fixed coverage, it has added resolution, and that, not the ECE reduction, is its contribution. If the ensemble's ECE holds steady across sub-benchmarks while a map fitted on one sub-benchmark fails on another, the ensemble buys robustness to shift, which is also a real contribution (Fact 5).

The case for the ensemble under shift. The third outcome deserves weight, and the telecom tables give some evidence for it. For the 4B model, raw ECE ranged from 0.147 to 0.800 across the three sub-benchmarks while the ensemble mean ranged from 0.093 to 0.108. A single map fitted on pooled data would have fitted the sub-benchmarks very differently. The ensemble's level was more stable across them, at least at this sample size. For the 12B model the ensemble range was 0.107 to 0.249 and for the 27B model 0.168 to 0.208, so the stability is less clear. Telecom deployments face exactly this problem, since questions about standards, specifications and code differ in difficulty and in how the model's confidence relates to them. Methods that need no labels avoid the cost of a fitted map for each domain, which is the motivation for cross-task calibrators such as Thermometer (Shen et al., 2024). The point of this section is not that ensembling is useless. It is that its claimed advantage over the cheap alternative is untested, and that the most plausible advantage, stability across domains, has to be tested across domains, which a pooled ECE does not do.

The right target for an auto-accept gate. A tier labelled "auto-accept" is a selective classifier. Its quality is the error among accepted answers at the coverage it accepts, and the curve of that error against coverage summarises the score across thresholds (Geifman and El-Yaniv, 2017; Traub et al., 2024). The threshold that minimises expected cost depends on the costs of a wrong acceptance and of a referral, as in Chow's rule (Chow, 1970), not on ECE. Where a guarantee is wanted, conformal methods give one under exchangeability of a calibration set and the deployment items. Conformal prediction sets for multiple-choice question answering with language models have been built and checked for coverage (Kumar et al., 2023). Distribution-free risk control picks a threshold so that the expected loss, for instance the error rate among accepted answers, stays below a target with high probability (Bates et al., 2021; Angelopoulos et al., 2022; Angelopoulos and Bates, 2021). KnowNo uses conformal prediction so that an LLM planner asks for help with a stated success guarantee (Ren et al., 2023). A threshold map for auto-

acceptance is exactly the object these methods certify. A guarantee of this kind holds only for items exchangeable with the calibration set, which brings the cross-domain question back. But it replaces a tier accuracy of 73 percent that appears to be computed on the evaluation items, with no held-out split reported, by a statement about future error at a stated confidence level.

A caution about human readers. Where the score is shown to a person rather than thresholded, calibration matters again, but not simply. People's confidence in model answers is miscalibrated relative to the model's own confidence. Aligning the wording of explanations with the model's internal confidence narrowed both the calibration gap and the discrimination gap in human judgements (Steyvers et al., 2025). An exploratory study found that showing advice as more confident than it was could improve joint human-AI accuracy, depending on how people use the advice (Vodrahalli et al., 2022). When the score is shown to a person, the ECE of the score is one input to the outcome, not the outcome itself.

10. The evidence in a table and a figure

Table 1 collects the measured values the argument rests on, one row per published result. ECE values are on each source's scale (Xiong et al. report them multiplied by 100). "Discrimination" is AUROC unless marked; Tian et al. (2023) report selective accuracy-coverage AUC, and Shrivastava et al. (2023) report it alongside AUROC. The two Saenko et al. (2026) tier rows are the author's arithmetic from that paper's Tables I to IV.

Figure 1(b) shows the dissociation directly. The largest ECE reductions (points 8 and 10, and the Twin-Pass ticks) have small or unknown discrimination changes. The largest discrimination gain (point 2) comes with a small ECE change. Point 4 got worse on ECE while its AUROC rose slightly, and point 11 got worse on both. If ECE and discrimination tracked each other, the points would lie along a falling diagonal from upper left to lower right. They do not. Figure 1(a) shows the sample-size problem of Section 8: on the 150-item rows, the ensemble ECE sits close to the noise tick; on the large 12B rows it sits well above it.

11. A decision procedure

Algorithm 1 states the decision structure the paper argues for: how to read a claim that a method improves LLM confidence, and what each branch licenses. It is a synthesis of the arguments above, not a procedure any cited study ran.

Applied to the telecom study, S1 classifies the use as a gate, since the paper proposes an auto-accept tier. S2 passes, since the answer is held fixed. S3 fails: no ranking metric is reported, and the claim of discriminative power is unsupported by the metrics given. S4 fails: no recalibrated single-pass baseline is reported, although the labels to fit one were in hand. S5 flags the 150-item rows as too small to separate the reported ECE from zero, and flags the 0.800 raw value as

not reconstructable. S6 is partly informative, since the ensemble's level was more stable than the raw score's across the 4B model's sub-benchmarks. S7 fails: the tier accuracies appear to be computed on the evaluation items, with no held-out split reported, and they have no stated coverage and carry no bound. The procedure does not overturn the result. It places it as evidence of a level improvement that may also include an unmeasured resolution improvement and some robustness across sub-benchmarks.

Applied to Tian et al. (2023), S3 notes that the reported AUC is selective, S4 passes because temperature-scaled ECE is reported for every method, and the conclusion that verbalised confidence beats label probability on level survives recalibration in their tables. Applied to Xiong et al. (2023), S3 passes, and their own reading, that prompting improves level while sampling improves ranking, is what the procedure would conclude.

12. Tests that would settle the open part

Five tests would separate what added inference buys from what recalibration buys. Each can be run on existing benchmarks with existing models.

T1. Add the ranking metrics to the existing comparison. For each model and sub-benchmark in the telecom study, compute AUROC and the error-coverage curve for the raw single pass, the ensemble mean and the ensemble median, with bootstrap intervals. The prediction that distinguishes the readings: if the ensemble mainly corrects level, its AUROC will be close to the single pass's. If it adds resolution, AUROC and error at a fixed coverage will both improve.

T2. Recalibrated single pass against ensemble, on held-out data. Split each sub-benchmark into a calibration half and an evaluation half. Fit histogram binning and isotonic regression to the single-pass score on the calibration half, and compare ECE with a debiased estimator (Roelofs et al., 2022; Kumar et al., 2019), AUROC and error at fixed coverage on the evaluation half, against the ensemble mean evaluated on the same half. Report the inference cost of each.

T3. Leave-one-domain-out transfer. Fit the single-pass map on two sub-benchmarks and evaluate on the third, for all three rotations, and do the same for any map applied to the ensemble mean. This tests directly the one advantage the ensemble most plausibly has, stability of level across domains (Fact 5, Section 9). If the unrecalibrated ensemble mean beats the transferred single-pass map on the held-out domain, the method's case is made on the ground that matters for deployment.

T4. Separate the passes from the averaging. Report the ensemble at $N = 1, 2, 4$ and 8 evaluation passes, and a cross-model evaluator (a different model family rating the same reasoning) at the same N . $N = 1$ isolates the effect of a separate evaluation call from the effect of averaging. Averaging adds granularity, and granularity matters for ranking because it breaks ties. The cross-model arm tests self-preference (Panickssery et al., 2024)

Table 1. Published changes in ECE and in discrimination for interventions on LLM confidence, one row per result, with the source on the row.

Source	Model, data	Intervention	ECE	Discrimination
Saenko et al. 2026	Gemma-3-4B, ORANBench (150)	single pass to twin-pass mean	0.800 to 0.093	not reported
Saenko et al. 2026	Gemma-3-4B, TeleQnA (1,000)	single pass to twin-pass mean	0.400 to 0.108	not reported
Saenko et al. 2026	Gemma-3-12B, TeleQnA (10,000)	single pass to twin-pass mean	0.314 to 0.244	not reported
Saenko et al. 2026	Gemma-3-12B, pooled (13,002)	top tier vs all answers	-	accuracy 0.73 vs 0.698
Saenko et al. 2026	Gemma-3-27B, pooled (1,300)	top tier vs all answers	-	accuracy 0.78 vs 0.745
Xiong et al. 2023	GPT-4, 8 datasets	vanilla verbalised	18.0	62.7
Xiong et al. 2023	GPT-4, GSM8K	CoT, all answers at 100%	6.4	constant score
Xiong et al. 2023	GPT-3.5, 5 datasets	1 CoT answer to 5 samples	25.0 to 18.7	56.4 to 73.0
Xiong et al. 2023	GPT-3.5, GSM8K	1 CoT answer to 5 samples	10.1 to 6.28	54.8 to 92.7
Xiong et al. 2023	GPT-3.5, StrategyQA	1 CoT answer to 5 samples	22.0 to 23.3	59.8 to 60.8
Xiong et al. 2023	GPT-4, 6 datasets	consistency to Pair-Rank	12.0 to 6.9	66.9 to 67.6
Tian et al. 2023	gpt-3.5, SciQ	label prob. to verbalised top-4	0.256 to 0.065	0.752 to 0.763 (sel.)
Tian et al. 2023	gpt-3.5, TriviaQA	label prob. to verbalised top-4	0.140 to 0.054	0.869 to 0.896 (sel.)
Tian et al. 2023	gpt-3.5, TriviaQA	1-stage top-1 to 2-stage CoT	0.068 to 0.110	0.879 to 0.830 (sel.)
Yoon et al. 2025	GLM-4 to GLM-Z1, NonambigQA	non-reasoning to reasoning	-0.037	+0.078
Podolak and Verma 2025	DeepSeek R1-32B	0 to 100-500 reasoning tokens	-	0.56 to 0.80
Chen and Mueller 2023	GPT-3.5, CommonsenseQA	self-reflection alone; combined	-	0.506; 0.769
Chen and Mueller 2023	GPT-3.5, GSM8K	self-reflection alone; combined	-	0.831; 0.951
Lyu et al. 2024	GPT-4, 4 domains	verbalised to consistency (FSD)	Brier 0.154 to 0.114	-
Shrivastava et al. 2023	GPT-4, 12 datasets	linguistic to mixture	-	AUROC 64.6 to 69.9 (selective AUC 80.5 to 83.4)
Kadavath et al. 2022	52B P(IK), LAMBADA	TriviaQA-only to all-task training	Brier 0.431 to 0.108	0.606 to 0.853
F. Zhu et al. 2023	ResNet110, CIFAR-10	baseline to label smoothing	calibration method	90.76 to 81.89

Algorithm 1. Reading a claim that a method improves LLM confidence

Input: a study P reporting score S_{new} against baseline S_{old}

S1: Name the use. Probability read by a person or a cost rule
 -> reliability matters. Threshold or gate -> ranking and error at coverage matter. If unstated, assume the gate.

S2: Did the answers change? If accuracy differs between arms, ECE changes may reflect accuracy moving toward a fixed level. Compare only score-side changes (S_{old} on A_{new}).

S3: Is a ranking metric reported (AUROC, error at coverage, AUC)? If no: P supports a calibration claim only. Check that "AUC" is ROC and not selective AUC.

S4: Is a recalibrated baseline reported? Fit a monotone map or histogram bins for S_{old} on held-out labels. If S_{new} does not beat recalibrated S_{old} on ECE, its ECE gain is a level effect a map buys at no inference cost.

S5: Is the ECE estimate sound? Check n per arm, binning scheme and bin count, and whether maps or thresholds were fitted on the evaluation items. On n of a few hundred, treat ECE below about 0.1 as indistinguishable from zero.

S6: Does the gain hold under shift? Fit on one domain, test on another; compare S_{new} 's stability with the map's.

S7: For a gate, report error among accepted answers at stated coverage, with a risk-control or conformal bound on a held-out calibration set.

Output: which property improved (level, resolution, robustness), at what cost, and how far it can be trusted.

directly in this setting, building on the paper's own 4B-critiques-12B observation.

T5. A certified gate. Using a held-out calibration set, choose the auto-accept threshold by distribution-free risk control so that the error rate among accepted answers is below a stated target with stated probability (Bates et al., 2021; Angelopoulos et al., 2022). Report the coverage each score achieves at that target. A score with better resolution will admit more answers at the same certified error rate, which is the operational quantity an engineer deciding whether to bypass review needs.

13. The strongest case against this paper

A reader may object on four grounds.

The first is that this is old news: calibration and discrimination have been distinguished since Murphy (1973) and DeGroot and Fienberg (1983), and the machine-learning failure-detection literature has made the same point recently (Jaeger et al., 2022; F. Zhu et al., 2023). That is correct, and the paper says so. The contribution is not the distinction. It is the demonstration that the language-model confidence literature, including a 2026 domain study, continues to support claims about discrimination with calibration metrics. It also lies in the consolidation of the cases where both were measured (Table 1), and in the specific re-reading of a recent result's own numbers. A reader who already applies Algorithm 1 to every confidence paper loses little.

The second is that ECE is what practitioners need, because operators want scores they can read as probabilities, and a well-calibrated score at the base rate is no worse than an honest admission of ignorance. The paper accepts this for the probability-reading use, and Section 9 notes that calibration matters when a person reads the score. But the telecom study proposes a gate, and for a gate a calibrated score with no resolution is useless: it passes everything or nothing. The objection also overlooks Fact 3. If calibration is the goal, a fitted map achieves it in distribution at no inference cost. The question is then whether extra passes beat the map, which the objection does not address.

The third is that recalibration needs labels, and labels are scarce in a new domain, so a method that improves calibration without labels is valuable even if a fitted map could match it in distribution. This is the strongest objection, and the paper partly concedes it (Section 9). Two replies. First, the telecom study used labels to compute ECE and to build its threshold map, so labels were available in the evaluated setting. Second, the advantage of not needing labels is an advantage under shift, and it must be tested under shift (T3). Pooled in-distribution ECE cannot show it.

The fourth is that the twin-pass method surely adds resolution, given the cross-model result, the tier ordering and the self-evaluation literature. Perhaps it does. The tier accuracies show some ranking, and the self-evaluation results in Section 7 are mixed rather than

negative. If T1 shows a large AUROC gain, this paper's reading of that study as evidence of a level improvement will have been too cautious. The general argument stands either way: the published evidence did not show it, and the metric used could not have.

14. What is not known

It is not known whether averaging separate-pass self-evaluations improves the resolution of confidence on multiple-choice technical questions. No study located here reports AUROC or error at coverage for such a method in the telecom domain. The one located decomposition of separate-pass self-reflection on a multiple-choice dataset found it at chance for GPT-3.5 (Chen and Mueller, 2023).

It is not known how much of any published ECE reduction for an inference-time confidence method would survive comparison with a recalibrated baseline fitted on held-out labels. Tian et al. (2023) report temperature-scaled ECE, and most other studies located here do not.

It is not known whether label-free confidence methods are more stable than fitted maps across domains within a technical field. The telecom tables are suggestive for one model size and inconclusive for the others, and no study located here runs a leave-one-domain-out comparison of the two.

It is not known whether the direction of the evaluation-generation gap, whether a model judges its own reasoning better or worse than it produces it, is a stable property of a model or varies by domain. Published results point both ways (Kadavath et al., 2022; Weng et al., 2023; West et al., 2023; Tyen et al., 2024; Stechly et al., 2024).

It is not known why the reported raw ECE of 0.800 for the 4B model on ORANBench is as large as it is (Section 8).

15. Limitations

The facts in Section 3 are elementary, and the numerical illustrations of estimator noise use a normal approximation with an assumed bin count and assume the confidences within a bin equal the bin's true accuracy. Real binned ECE at small n depends on the binning scheme, on how many distinct values the score takes and on the distribution of accuracy across bins, so the noise ticks in Figure 1(a) are indicative, not estimates for the telecom study's actual procedure.

The pooled accuracies in Section 8 assume the threshold map pools the three sub-benchmarks with the item counts in Tables I to III. If it was built on a different subset, the comparison of top-tier and base-rate accuracy changes, though the absence of tier shares and of a ranking metric does not.

Table 1 and Figure 1(b) put together results that use different models, datasets, numbers of samples and metrics. One of the sources (Tian et al., 2023) reports

only selective AUC, and the figure marks its points separately; Shrivastava et al. (2023) report both. The ECE changes are on each source's scale and binning. The figure shows the pattern of dissociation, not a pooled effect size, and no regression line is drawn because none would be meaningful.

The collection is not exhaustive. It includes the studies located that report both a calibration and a ranking metric for the same intervention, and studies cited for the facts they establish. Studies reporting only ECE are under-represented in Table 1 by design, since they cannot contribute to the dissociation question. The image-classification results of Section 6 come from a different setting and are used only to show that calibration methods can damage ranking, not to estimate how often that happens for language models.

The paper re-reads one 2026 telecom study in detail because it is a clear recent case of the pattern, not because it is unusual. Several of its gaps (no ranking metric, no recalibrated baseline, no held-out split reported for its thresholds) are common in the literature.

16. Conclusion

"The model's confidence was overconfident, and our method fixed it" is two claims. Overconfidence as a level problem is widespread, well documented and cheap to correct in distribution: a map fitted on a few hundred labelled answers removes most of it and changes no ranking. Overconfidence as a resolution problem, scores that do not separate right answers from wrong ones, is documented wherever a ranking metric has been computed. It cannot be fixed by any function of the score alone. Sampling, separate evaluation passes and longer reasoning can add resolution. Answer sampling demonstrably does, by amounts that range from negligible to large across datasets; longer reasoning has done so in the two studies that measured it; and for a separate self-evaluation pass of a fixed answer the evidence is mixed, with one located measurement on a multiple-choice set at chance. ECE measures the first problem and is blind to the second, and in the studies that measured both, the two moved independently.

The telecom twin-pass study shows a method that reduced calibration error, most dramatically on one 150-item subset, and that may also have added resolution it did not measure. Its auto-accept tier was about three points more accurate than accepting everything. Whether its extra passes beat a recalibrated single pass, whether they add ranking, and whether their level holds across telecom subdomains better than a fitted map are open questions, and each has a direct test. Until such tests are run, a fall in ECE should be read as what it is: evidence that the scores are closer to the right level, not evidence that they know which answers are wrong.

References

- Angelopoulos, A. N., & Bates, S. (2021). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511. doi:10.48550/arXiv.2107.07511.
- Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2022). Conformal Risk Control. arXiv:2208.02814. doi:10.48550/arXiv.2208.02814.
- Band, N., Li, X., Ma, T., & Hashimoto, T. (2024). Linguistic Calibration of Long-Form Generations. arXiv:2404.00474. doi:10.48550/arXiv.2404.00474.
- Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. I. (2021). Distribution-Free, Risk-Controlling Prediction Sets. arXiv:2101.02703. doi:10.48550/arXiv.2101.02703.
- Blasiok, J., Gopalan, P., Hu, L., & Nakkiran, P. (2023). A Unifying Theory of Distance from Calibration. arXiv:2211.16886. doi:10.48550/arXiv.2211.16886.
- Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1), 1-3.
- Chen, J., & Mueller, J. (2023). Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness. arXiv:2308.16175. doi:10.48550/arXiv.2308.16175.
- Chow, C. (1970). On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1), 41-46. doi:10.1109/TIT.1970.1054406
- Damani, M., Puri, I., Slocum, S., Shenfeld, I., Choshen, L., Kim, Y., & Andreas, J. (2025). Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. arXiv:2507.16806. doi:10.48550/arXiv.2507.16806.
- DeGroot, M. H., & Fienberg, S. E. (1983). The Comparison and Evaluation of Forecasters. *The Statistician*, 32(1/2), 12-22. doi:10.2307/2987588
- Desai, S., & Durrett, G. (2020). Calibration of Pre-trained Transformers. arXiv:2003.07892. doi:10.48550/arXiv.2003.07892.
- Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., et al. (2023). LM-Polygraph: Uncertainty Estimation for Language Models. arXiv:2311.07383. doi:10.48550/arXiv.2311.07383.
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625-630. doi:10.1038/s41586-024-07421-0
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010
- Gajjar, P., & Shah, V. K. (2024). ORAN-Bench-13K: An Open Source Benchmark for Assessing LLMs in Open Radio Access Networks. arXiv:2407.06245. doi:10.48550/arXiv.2407.06245.
- Geifman, Y., & El-Yaniv, R. (2017). Selective Classification for Deep Neural Networks. arXiv:1705.08500. doi:10.48550/arXiv.1705.08500.
- Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., et al. (2025). Gemma 3 Technical Report. arXiv:2503.19786. doi:10.48550/arXiv.2503.19786.
- Geng, J., Cai, F., Wang, Y., Koepl, H., Nakov, P., & Gurevych, I. (2024). A Survey of Confidence Estimation and Calibration in Large Language Models. arXiv:2311.08298. doi:10.48550/arXiv.2311.08298.
- Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2), 243-268. doi:10.1111/j.1467-9868.2007.00587.x
- Groot, T., & Valdenegro-Toro, M. (2024). Overconfidence is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models. arXiv:2405.02917. doi:10.48550/arXiv.2405.02917.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. arXiv:1706.04599. doi:10.48550/arXiv.1706.04599.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.

- Radiology, 143(1), 29-36. doi:10.1148/radiology.143.1.7063747
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. arXiv:2310.01798. doi:10.48550/arXiv.2310.01798.
- Jaeger, P. F., Luth, C. T., Klein, L., & Bungert, T. J. (2022). A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification. arXiv:2211.15259. doi:10.48550/arXiv.2211.15259.
- Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. arXiv:2012.00955. doi:10.48550/arXiv.2012.00955.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221. doi:10.48550/arXiv.2207.05221.
- Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., et al. (2024). Large Language Models Must Be Taught to Know What They Don't Know. arXiv:2406.08391. doi:10.48550/arXiv.2406.08391.
- Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., & Gal, Y. (2024). Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. arXiv:2406.15927. doi:10.48550/arXiv.2406.15927.
- Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. arXiv:2302.09664. doi:10.48550/arXiv.2302.09664.
- Kumar, A., Liang, P., & Ma, T. (2019). Verified Uncertainty Calibration. arXiv:1909.10155. doi:10.48550/arXiv.1909.10155.
- Kumar, A., Morabito, R., Umbet, S., Kabbara, J., & Emami, A. (2024). Confidence Under the Hood: An Investigation into the Confidence-Probability Alignment in Large Language Models. arXiv:2405.16282. doi:10.48550/arXiv.2405.16282.
- Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., & Beam, A. (2023). Conformal Prediction with Large Language Models for Multi-Choice Question Answering. arXiv:2305.18404. doi:10.48550/arXiv.2305.18404.
- Leng, J., Huang, C., Zhu, B., & Huang, J. (2024). Taming Overconfidence in LLMs: Reward Calibration in RLHF. arXiv:2410.09724. doi:10.48550/arXiv.2410.09724.
- Lin, S., Hilton, J., & Evans, O. (2022). Teaching Models to Express Their Uncertainty in Words. arXiv:2205.14334. doi:10.48550/arXiv.2205.14334.
- Lyu, Q., Shridhar, K., Malaviya, C., Zhang, L., Elazar, Y., Tandon, N., et al. (2024). Calibrating Large Language Models with Sample Consistency. arXiv:2402.13904. doi:10.48550/arXiv.2402.13904.
- Maatouk, A., Ayed, F., Piovesan, N., De Domenico, A., Debbah, M., & Luo, Z. (2023). TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge. arXiv:2310.15051. doi:10.48550/arXiv.2310.15051.
- Mahaut, M., Aina, L., Czarnowska, P., Hardalov, M., Muller, T., & Marquez, L. (2024). Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators. arXiv:2406.13415. doi:10.48550/arXiv.2406.13415.
- Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896. doi:10.48550/arXiv.2303.08896.
- Mei, Z., Zhang, C., Yin, T., Lidard, J., Shorinwa, O., & Majumdar, A. (2025). Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know?. arXiv:2506.18183. doi:10.48550/arXiv.2506.18183.
- Mielke, S. J., Szlam, A., Dinan, E., & Boureau, Y. (2022). Reducing conversational agents' overconfidence through linguistic calibration. arXiv:2012.14983. doi:10.48550/arXiv.2012.14983.
- Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12(4), 595-600.
- Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). doi:10.1609/aaai.v29i1.9602
- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625-632. doi:10.1145/1102351.1102430
- Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., & Tran, D. (2019). Measuring Calibration in Deep Learning. arXiv:1904.01685. doi:10.48550/arXiv.1904.01685.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2023). GPT-4 Technical Report. arXiv:2303.08774. doi:10.48550/arXiv.2303.08774.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., et al. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. arXiv:1906.02530. doi:10.48550/arXiv.1906.02530.
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076. doi:10.48550/arXiv.2404.13076.
- Podolak, J., & Verma, R. (2025). Read Your Own Mind: Reasoning Helps Surface Self-Confidence Signals in LLMs. arXiv:2505.23845. doi:10.48550/arXiv.2505.23845.
- Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., et al. (2023). Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. arXiv:2307.01928. doi:10.48550/arXiv.2307.01928.
- Roelofs, R., Cain, N., Shlens, J., & Mozer, M. C. (2022). Mitigating Bias in Calibration Error Estimation. arXiv:2012.08668. doi:10.48550/arXiv.2012.08668.
- Saenko, A., Gajjar, P., Ganiyu, A., & Shah, V. K. (2026). Enhancing Confidence Estimation in Telco LLMs via Twin-Pass CoT-Ensembling. arXiv:2604.13271. doi:10.48550/arXiv.2604.13271.
- Shen, M., Das, S., Greenewald, K., Sattigeri, P., Wornell, G., & Ghosh, S. (2024). Thermometer: Towards Universal Calibration for Large Language Models. arXiv:2403.08819. doi:10.48550/arXiv.2403.08819.
- Shrivastava, V., Liang, P., & Kumar, A. (2023). Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation. arXiv:2311.08877. doi:10.48550/arXiv.2311.08877.
- Stechly, K., Valmeekam, K., & Kambhampati, S. (2024). On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. arXiv:2402.08115. doi:10.48550/arXiv.2402.08115.
- Stengel-Eskin, E., Hase, P., & Bansal, M. (2024). LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models. arXiv:2405.21028. doi:10.48550/arXiv.2405.21028.
- Steyvers, M., Tejada, H., Kumar, A., Belem, C., Karny, S., Hu, X., et al. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221-231. doi:10.1038/s42256-024-00976-7
- Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., et al. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. arXiv:2305.14975. doi:10.48550/arXiv.2305.14975.
- Traub, J., Bungert, T. J., Luth, C. T., Baumgartner, M., Maier-Hein, K. H., Maier-Hein, L., & Jaeger, P. F. (2024). Overcoming Common Flaws in the Evaluation of Selective Classification Systems. arXiv:2407.01032. doi:10.48550/arXiv.2407.01032.
- Tyen, G., Mansoor, H., Carbune, V., Chen, P., & Mak, T. (2024). LLMs cannot find reasoning errors, but can correct them given the error location. arXiv:2311.08516. doi:10.48550/arXiv.2311.08516.
- Ulmer, D., Gubri, M., Lee, H., Yun, S., & Oh, S. J. (2024). Calibrating Large Language Models Using Their Generations Only. arXiv:2403.05973. doi:10.48550/arXiv.2403.05973.

- Vashurin, R., Fadeeva, E., Vazhentsev, A., Rvanova, L., Tsvigun, A., Vasilev, D., et al. (2024). Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. arXiv:2406.15627. doi:10.48550/arXiv.2406.15627.
- Vodrahalli, K., Gerstenberg, T., & Zou, J. (2022). Uncalibrated Models Can Improve Human-AI Collaboration. arXiv:2202.05983. doi:10.48550/arXiv.2202.05983.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., et al. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171. doi:10.48550/arXiv.2203.11171.
- Wataoka, K., Takahashi, T., & Ri, R. (2024). Self-Preference Bias in LLM-as-a-Judge. arXiv:2410.21819. doi:10.48550/arXiv.2410.21819.
- Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., et al. (2023). Large Language Models are Better Reasoners with Self-Verification. arXiv:2212.09561. doi:10.48550/arXiv.2212.09561.
- West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., et al. (2023). The Generative AI Paradox: "What It Can Create, It May Not Understand". arXiv:2311.00059. doi:10.48550/arXiv.2311.00059.
- Xia, Y., Ulmer, D., Blevins, T., Liu, Y., Schutze, H., & Roth, B. (2026). Calibration Is Not Enough: Evaluating Confidence Estimation Under Language Variations. arXiv:2601.08064. doi:10.48550/arXiv.2601.08064.
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. arXiv:2306.13063. doi:10.48550/arXiv.2306.13063.
- Xu, T., Wu, S., Diao, S., Liu, X., Wang, X., Chen, Y., & Gao, J. (2024). SaySelf: Teaching LLMs to Express Confidence with Self-Reflective Rationales. arXiv:2405.20974. doi:10.48550/arXiv.2405.20974.
- Yang, D., Tsai, Y. H., & Yamada, M. (2024). On Verbalized Confidence Scores for LLMs. arXiv:2412.14737. doi:10.48550/arXiv.2412.14737.
- Yona, G., Aharoni, R., & Geva, M. (2024). Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?. arXiv:2405.16908. doi:10.48550/arXiv.2405.16908.
- Yoon, D., Kim, S., Yang, S., Kim, S., Kim, S., Kim, Y., et al. (2025). Reasoning Models Better Express Their Confidence. arXiv:2505.14489. doi:10.48550/arXiv.2505.14489.
- Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 694-699. doi:10.1145/775047.775151
- Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685. doi:10.48550/arXiv.2306.05685.
- Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2024). Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. arXiv:2401.06730. doi:10.48550/arXiv.2401.06730.
- Zhou, K., Jurafsky, D., & Hashimoto, T. (2023). Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. arXiv:2302.13439. doi:10.48550/arXiv.2302.13439.
- Zhu, C., Xu, B., Wang, Q., Zhang, Y., & Mao, Z. (2023). On the Calibration of Large Language Models and Alignment. arXiv:2311.13240. doi:10.48550/arXiv.2311.13240.
- Zhu, F., Cheng, Z., Zhang, X., & Liu, C. (2023). Rethinking Confidence Calibration for Failure Prediction. arXiv:2303.02970. doi:10.48550/arXiv.2303.02970.