Beyond F1: A Cross-Model Calibration Audit for Extractive Question Answering

Anonymous ACL submission

Abstract

When a question answering system says it is confident, should you believe it? Extractive QA models, which answer questions by identifying a span of text within a passage, are widely deployed in real-world applications yet evaluated almost exclusively on F1 score. This tells us whether a model is accurate, but nothing about whether its confidence signals are trustworthy. A model can achieve a high F1 and remain dangerously miscalibrated, confidently wrong in systematic and predictable ways that vary by question type, answer length, and domain. Critically, this problem has never been studied comparatively across model families. In this paper, we present **Calibrated-QA-Bench**, the first cross-model, subgroup-aware calibration audit for extractive QA. Across three model families evaluated on SQuAD 2.0 and CUAD, we show that miscalibration is not random noise but a structured phenomenon, with certain question types consistently twice as miscalibrated as others, regardless of the model used. Extending to a generative LLM baseline, we uncover a striking contrast: while LLMs are better calibrated when an answer exists, they almost entirely fail to recognize when they should not answer at all, a failure mode opposite to that of extractive models. We further show that asking an LLM to verbally rate its own confidence reveals this uncertainty far more reliably than its internal probability scores alone. Finally, we train a lightweight calibration corrector and show that calibration patterns learned on one model family transfer to a held-out architecture, reducing miscalibration by over 90%. Code and benchmarks are publicly available at `https://tinyurl.com/qa-bench-aacl`.

1 Introduction

Extractive question answering (QA), the task of identifying an answer span within a provided passage, has been a cornerstone of NLP research since the introduction of large-scale reading comprehension benchmarks (Rajpurkar et al., 2016, 2018). While the dominance of large language models (LLMs) has shifted research attention toward generative approaches, extractive QA systems remain widely deployed in practice. In latency-sensitive, resource-constrained, or privacy-critical environments such as clinical decision support, legal document review, and enterprise search, lightweight fine-tuned models offer deterministic inference, interpretable grounded outputs, and compliance with data governance requirements that LLM-based systems cannot always satisfy (Sung et al., 2021; Manor and Li, 2019). Furthermore, as trust and reliability of LLM outputs become central concerns (Xiong et al., 2024), understanding the confidence calibration of extractive models provides an important reference point for evaluating whether newer generative systems represent a genuine improvement in uncertainty estimation.

Despite this continued relevance, a critical limitation of extractive QA evaluation has gone largely unaddressed. Models are assessed almost exclusively on F1 score and Exact Match, metrics that measure prediction accuracy but say nothing about whether a model's confidence in its predictions is reliable. Calibration, the property that a model's stated confidence should match its empirical accuracy, has been studied extensively in image classification (Guo et al., 2017) and more recently in natural language generation (Xiong et al., 2024), but remains poorly understood in extractive QA. A model that scores 85% F1 can simultaneously be dangerously overconfident, assigning near-certain confidence to predictions it frequently gets wrong. In safety-critical deployments, this miscalibration is not a theoretical concern but a practical failure: an overconfident wrong answer is actively more harmful than an abstention (Kamath et al., 2020).

What makes this gap particularly significant is that calibration has never been studied comparatively across extractive QA model families. Prior work on QA calibration has examined single mod-

els in isolation (Kamath et al., 2020; Varshney et al., 2022), leaving open whether miscalibration varies systematically by subgroup, whether architectural choices affect calibration patterns, and whether LLMs produce better-calibrated uncertainty signals than fine-tuned extractive models. These questions have direct implications for how practitioners should select and deploy QA systems, yet no existing benchmark provides the infrastructure to answer them.

In this paper, we present **Calibrated-QA-Bench**, the first systematic, cross-model, subgroup-aware calibration audit framework for extractive QA. We evaluate three extractive model families, RoBERTa-base, RoBERTa-large, and DeBERTa-v3-base, on SQuAD 2.0 (Rajpurkar et al., 2018) and CUAD (Hendrycks et al., 2021), measuring Expected Calibration Error (ECE) (Guo et al., 2017) and selective prediction quality across fine-grained subgroups defined by question type, answer length, and answerability. We extend the audit to a generative LLM baseline, Llama-3.2-3B-Instruct, extracting both token-level logprob confidence and verbalized uncertainty scores. Finally, we train a lightweight cross-model calibration corrector that predicts answer correctness from confidence signals and subgroup features, demonstrating that calibration patterns learned on one model family generalise to a held-out architecture. Our key findings are as follows:

1. Miscalibration in extractive QA is **structured and predictable**, varying systematically by question type and answer length across all three model families, with architectural choices producing distinct **calibration fingerprints**: DeBERTa is dramatically better calibrated on unanswerable questions but worse on answerable ones compared to RoBERTa variants.

2. LLMs and extractive models exhibit **opposite calibration failure modes**: the LLM is well-calibrated when an answer exists but catastrophically miscalibrated on abstention decisions, while extractive models show the reverse pattern.

3. **Verbalized confidence functions as an abstention detector** rather than a graded signal, capturing uncertainty that logprob confidence fails to express; a lightweight cross-model calibration corrector exploiting this reduces miscalibration on a **held-out architecture by over 90%** without retraining.

2 Related Work

Confidence calibration in neural networks was systematically studied by Guo et al. (2017), who introduced Expected Calibration Error and demonstrated that modern neural networks are significantly overconfident, proposing temperature scaling as a post-hoc remedy. While calibration has since been studied in image classification and natural language generation, but its application to extractive QA has received limited attention. Kamath et al. (2020) showed that softmax-based confidence fails under domain shift in QA and proposed a calibrator to improve selective prediction, directly motivating our calibration corrector design. Varshney et al. (2022) extended this to multiple NLP tasks, establishing a formal risk-coverage framework for selective prediction that we adopt in our evaluation. However, both works examine single models in isolation, leaving no prior work that studies calibration comparatively across extractive QA model families, measures subgroup-level calibration patterns, or tests whether a calibration corrector trained on one architecture generalises to another.

The calibration of large language models has emerged as an active research area. Xiong et al. (2024) provides a comprehensive survey of confidence estimation methods in LLMs, covering token-level logprob confidence, verbalized uncertainty elicitation, and consistency-based sampling, noting that these signals capture different aspects of uncertainty and that their relative reliability across task types remains an open empirical question. Our work contributes direct empirical evidence on this question in the context of extractive QA: we show that verbalized confidence functions primarily as an abstention detector rather than a graded confidence signal, a finding with practical implications for how these signals should be combined in deployed systems. To our knowledge, no prior work has directly compared the calibration failure modes of extractive and generative models on the same benchmark under the same evaluation framework.

3 Datasets and Models

3.1 Datasets

We evaluate on two datasets that together provide both an in-domain calibration benchmark and a domain shift stress test.

SQuAD 2.0 (Rajpurkar et al., 2018) is the primary evaluation dataset. It contains 11,873 validation examples drawn from Wikipedia articles, with questions that are either answerable by a span in the passage or deliberately unanswerable. The validation split is nearly 50/50 answerable versus unanswerable, making it well suited for evaluating a model's ability to calibrate both its span predictions and its abstention decisions. All extractive models are evaluated on the full validation set. For the LLM comparison, we draw a stratified sample of 3,000 examples proportional by question type, taking all available Yes/No questions to ensure sufficient coverage of the smallest subgroup.

CUAD (Hendrycks et al., 2021) is used exclusively as a domain shift evaluation target. It contains 4,182 test examples drawn from 510 commercial legal contracts annotated by expert lawyers across 41 clause categories. Crucially, we do not fine-tune any model on CUAD. Instead, we evaluate SQuAD-trained extractive models directly on CUAD to measure how calibration degrades when a model encounters text that is structurally and stylistically unlike its training distribution. This tests whether confidence signals that appear reliable in-domain remain meaningful under domain shift, a practically important question for any deployed QA system.

Table 1 summarizes the dataset statistics.

Dataset	Split	Total	Ans.(%)	Unans.(%)
SQuAD 2.0	Train	130,319	66.6	33.4
SQuAD 2.0	Validation	11,873	49.9	50.1
CUAD	Train	22,450	49.8	50.2
CUAD	Test	4,182	29.7	70.3

Table 1: Dataset statistics. Ans. and Unans. refer to the percentage of answerable and unanswerable examples, respectively.

3.2 Models

We evaluate four models spanning three architectural families, designed to test the effect of model size within the same family (RoBERTa-base vs large), architectural choice across families (RoBERTa vs DeBERTa), and the generative paradigm versus fine-tuned extraction (Llama vs all extractive models). RoBERTa-base serves as the primary baseline. DeBERTa's disentangled attention mechanism differs fundamentally from RoBERTa's standard self-attention, making it a meaningful architectural comparison at comparable parameter scale. Llama-3.2-3B-Instruct is prompted to extract an answer span or respond with *unanswerable* if no answer exists. Table 2 summarises the model lineup.

Model	HuggingFace ID	Params
RoBERTa-base	roberta-base-squad2	125M
RoBERTa-large	roberta-large-squad2	355M
DeBERTa-v3	deberta-v3-base-squad2	184M
Llama-3.2-3B	Llama-3.2-3B-Instruct	3B

Table 2: Models evaluated (all via HuggingFace `deepset/` prefix for extractive models, `meta-llama/` for Llama). All extractive models evaluated in float32 without post-hoc calibration. Llama evaluated zero-shot in float16 on a P100 GPU.

4 Methodology

Figure 1 provides an overview of the Calibrated-QA-Bench framework. A QA example consisting of a question and a context passage is passed through either an extractive model or a generative LLM. Confidence signals are extracted from each model type, every prediction is tagged with subgroup labels, and the resulting annotated predictions are used for both calibration analysis and trust module training. We describe each component in detail below.

4.1 Confidence Signal Extraction

Extractive Models. For each input pair (question, context), an extractive QA model produces start logits $\mathbf{s} \in \mathbb{R}^L$ and end logits $\mathbf{e} \in \mathbb{R}^L$ over all L tokens in the encoded input. Following standard practice (Rajpurkar et al., 2016), the score of a candidate answer span (i, j) is computed as:

$$\text{score}(i, j) = s_i + e_j, \quad i \leq j \tag{1}$$

The predicted answer is the span maximising Equation 1 subject to a maximum span length constraint. We extract three signals from this computation for every prediction.

Confidence is the score of the best predicted span:

$$\text{conf} = \max_{i \leq j} (s_i + e_j) \tag{2}$$

Null score is the score of the null span at position $(0, 0)$, which represents the model's internal signal for "no answer exists in this passage":

$$\text{null} = s_0 + e_0 \tag{3}$$

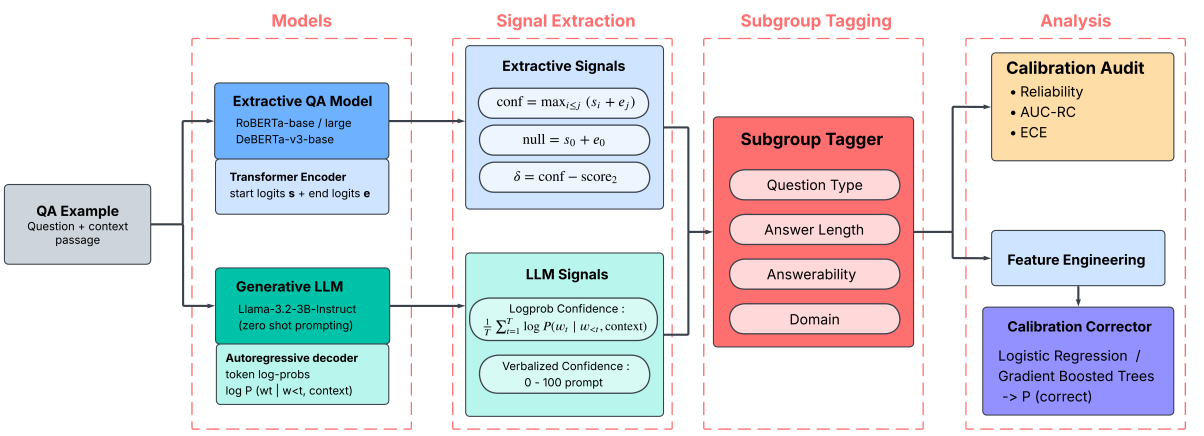

Figure 1: Overview of the Calibrated-QA-Bench framework. A QA example is passed through either an extractive model or a generative LLM. Confidence signals are extracted and tagged with subgroup labels before being used for calibration analysis and calibration corrector training.

Delta is the margin between the top-ranked and second-ranked span scores, measuring how decisive the model is in its prediction:

$$\delta = \text{conf} - \text{score}_2 \tag{4}$$

where score$_2$ is the score of the second-best candidate span. A small delta indicates the model was nearly indifferent between two candidate answers, which we expect to correlate with lower correctness.

Generative LLM. For the LLM, we extract two qualitatively different confidence signals. **Logprob confidence** is the mean log probability of generated tokens:

$$\text{conf}_{\text{lp}} = \frac{1}{T}\sum_{t=1}^{T} \log P(w_t \mid w_{<t}, \text{context}) \tag{5}$$

This is the closest analogue to the extractive span score in Equation 2, as both reflect the model's internal probability assigned to its own output. **Verbalized confidence** is obtained by prompting the model to rate its confidence on a scale of 0–100 after generating an answer, normalized to $[0, 1]$. As we show in Section 5.4, these two signals capture fundamentally different aspects of uncertainty and are complementary rather than redundant.

4.2 Subgroup Taxonomy

Every prediction is tagged across four subgroup axes. This allows calibration to be measured not only globally but within fine-grained, semantically meaningful slices of the data.

Question type is assigned from the first word of the question: *what* and *which* map to WHAT; *who* and *whose* to WHO; *is, are, was, were, did, does, do, can, could,* and *would* to YESNO; and *when, where, how,* and *why* to their respective categories. CUAD questions follow a template structure and are assigned a single CUADCLAUSE type.

Answer length is binned by the token count of the gold answer span: SHORT (1–3 tokens), MEDIUM (4–10 tokens), LONG (more than 10 tokens), and NONE for unanswerable questions.

Answerability is binary: ANSWERABLE if a gold answer span exists in the passage, UNANSWERABLE otherwise.

Domain distinguishes SQUAD examples from CUAD examples and enables direct measurement of calibration degradation under domain shift.

Table 3 shows the subgroup distribution across the SQuAD 2.0 validation set.

Axis	Label	Count
Question Type	What / Which	6,140
	Other	2,159
	How	1,131
	Who / Whose	1,018
	When	736
	Where	429
	Why	187
	Yes/No	60
Answer Length	None (unans.)	5,945
	Short	4,360
	Medium	1,309
	Long	259
Answerability	Unanswerable	5,945
	Answerable	5,928

Table 3: Subgroup distribution for the SQuAD 2.0 validation set across the three primary subgroup axes.

4.3 Calibration Metrics

Expected Calibration Error. A model is well-calibrated if its stated confidence matches its empirical accuracy: when the model assigns confidence c to a set of predictions, approximately c fraction of those predictions should be correct. Expected Calibration Error (ECE) (Guo et al., 2017) quantifies the deviation from this ideal by partitioning predictions into M equally-spaced confidence bins and computing the weighted average gap between mean confidence and mean accuracy within each bin:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \quad (6)$$

where B_m is the set of predictions whose confidence falls in bin m, N is the total number of predictions, $\text{acc}(B_m)$ is the fraction of correct predictions in the bin, and $\text{conf}(B_m)$ is the mean confidence in the bin. A perfectly calibrated model has ECE = 0. We use $M = 10$ bins throughout. Correctness is determined by a token-level F1 threshold of 0.5, consistent with standard SQuAD evaluation (Rajpurkar et al., 2016). For unanswerable questions, a prediction is correct if and only if the model abstains by predicting an empty span.

Reliability Diagrams. Reliability diagrams plot mean confidence against mean accuracy per bin. A perfectly calibrated model falls on the diagonal; deviations indicate overconfidence (below) or underconfidence (above).

Risk-Coverage Analysis. The risk-coverage curve evaluates confidence as an abstention signal. At each confidence threshold τ, we compute coverage (the fraction of examples for which the model produces an answer) and risk (the error rate on answered examples). Sweeping τ from 1 to 0 traces the full risk-coverage curve. The area under this curve (AUC-RC) summarises selective prediction quality, where lower values indicate that the model's confidence is more useful for identifying which predictions to trust (Varshney et al., 2022). Unlike ECE, which measures numeric calibration, AUC-RC measures the ranking quality of the confidence signal as a decision threshold.

4.4 Confidence Normalization Strategy

Raw span scores are unbounded real numbers with no natural probabilistic interpretation. For **single-model calibration analysis** we apply sigmoid normalization, consistent with published benchmarks (Kamath et al., 2020):

$$\hat{c} = \frac{1}{1 + e^{-\text{conf}}} \quad (7)$$

For **cross-model comparison** we apply min-max normalization, because RoBERTa-large produces logit magnitudes substantially larger than RoBERTa-base or DeBERTa, causing sigmoid to saturate and compress nearly all predictions to values near 1.0:

$$\hat{c} = \frac{\text{conf} - \min(C)}{\max(C) - \min(C) + \epsilon} \quad (8)$$

This means that ECE values in the single-model analysis (Sections 5.1 and 5.2) are not directly comparable to those in the cross-model comparison (Section 5.3). For example, RoBERTa-base yields ECE = 0.161 under sigmoid normalization and ECE = 0.373 under min-max normalization — both are correct within their respective contexts. LLM logprob confidence uses min-max normalization since mean log probabilities are negative reals in $(-\infty, 0]$; verbalized confidence requires no normalization as it is already in $[0, 1]$.

4.5 Calibration Corrector

The calibration corrector is a lightweight meta-classifier that replaces raw confidence as the downstream decision signal. Rather than relying on a single uncalibrated number, the corrector predicts $P(\text{correct})$ from a richer feature set combining confidence signals with subgroup context. We construct a 16-dimensional feature vector: the three continuous signals (confidence, null score, and delta) are each min-max normalized per the model's full prediction set; answerability is encoded as a binary feature; and question type and answer length are each one-hot encoded, yielding 8 and 4 features respectively.

Classifiers. We train two classifiers on this feature vector. Logistic Regression (LR) provides an interpretable baseline whose coefficients reveal the linear contribution of each feature to correctness prediction. Gradient Boosted Trees (GBT) provide a stronger non-linear classifier; including both tests whether the feature-correctness relationship is approximately linear, which is itself informative about the nature of the calibration signal. The default hyperparamters were used for both the models without any fine-tuning.

Cross-Model Train and Test Split. The key design choice is a cross-model train and test split. We train exclusively on the combined predictions of RoBERTa-base and RoBERTa-large on SQuAD 2.0 (23,746 training examples), then test on DeBERTa-v3-base predictions held out entirely during training. DeBERTa belongs to a different architectural family with a different attention mechanism and different logit scale. If the corrector improves calibration on DeBERTa despite never having seen its predictions, it demonstrates that confidence signals, subgroup features, and answer correctness generalise across model families — providing evidence for a model-agnostic trust layer any extractive QA system can adopt without retraining.

5 Results

5.1 Baseline Calibration Audit

Table 4 reports global F1, ECE, and AUC-RC for all four models on SQuAD 2.0. RoBERTa-base achieves 82.5% F1, consistent with the reported benchmark for this model (Rajpurkar et al., 2018), confirming that our inference pipeline is correctly implemented.

Model	F1	ECE	AUC-RC
Extractive QA Models			
RoBERTa-base	82.5	0.373	0.112
RoBERTa-large	88.3	0.417	**0.060**
DeBERTa-v3-base	87.3	**0.342**	0.073
Generative LLM			
Llama-3.2-3B (3k)	54.2	0.343	—

Table 4: Global F1, ECE, and AUC-RC on SQuAD 2.0. Lower ECE and AUC-RC indicate better calibration. **Bold** indicates best value per column. AUC-RC is not reported for Llama as risk-coverage analysis requires logit-based span scores.

Despite achieving the highest F1 score, RoBERTa-large has the worst global ECE (0.417), demonstrating that accuracy and calibration are not the same property. DeBERTa achieves the best global ECE (0.342) while also achieving strong F1 (87.3%). The Llama baseline achieves comparable global ECE (0.343) to the extractive models despite receiving no fine-tuning on SQuAD, a result that requires careful interpretation as we discuss in Section 5.4. RoBERTa-base exemplifies the defining calibration failure of extractive QA models: it assigns a mean confidence of 0.987 across all predictions while achieving only 82.7% accuracy — a 16-point overconfidence gap — with nearly all predictions compressed above 0.95 regardless of correctness. This confidence compression renders the raw signal useless as a graded decision threshold, motivating the subgroup and cross-model analysis that follows.

5.2 Miscalibration Is Structured

Figure 3 shows reliability diagrams broken down by question type for RoBERTa-base. Miscalibration varies substantially across subgroups: Yes/No questions exhibit an ECE of 0.295 while How questions exhibit an ECE of 0.150, a nearly 2× gap. Why questions (ECE = 0.236) and long-answer questions (ECE = 0.227) are also substantially worse calibrated than the global average. Critically, this pattern is not model-specific. Figure 4 shows that the same subgroup ranking holds across all three extractive model families: Yes/No questions are consistently the most miscalibrated type and How questions are consistently the best calibrated, regardless of model size or architecture.

The risk-coverage analysis reinforces this finding. Yes/No questions reach only 60% coverage at a 20% risk tolerance, while How questions reach 100% coverage at the same threshold. This 40-point coverage gap proves that a single global confidence threshold cannot serve all subgroups equally and directly motivates per-subgroup calibration intervention.

5.3 Cross-Model Calibration Comparison

Table 5 reveals the most consequential architectural difference across models. DeBERTa is dramatically better calibrated on unanswerable questions (ECE = 0.067) than RoBERTa-base (ECE = 0.302) but dramatically worse on answerable questions (ECE = 0.670 vs 0.446). This split personality, where the same model is simultaneously the best and the worst calibrated depending on answerability, suggests that DeBERTa's disentangled attention mechanism produces a stronger abstention signal at the cost of overconfidence on span predictions. RoBERTa-large, despite higher accuracy, produces the worst answerable ECE (0.517), reinforcing that model size does not reliably improve calibration.

5.4 LLM Calibration and Verbalized Confidence

Table 5 already reveals the headline finding for the LLM comparison. While Llama achieves the best answerable ECE of any studied model (0.080), it catastrophically fails on unanswerable questions with ECE = 0.615, achieving only 25.2% accuracy

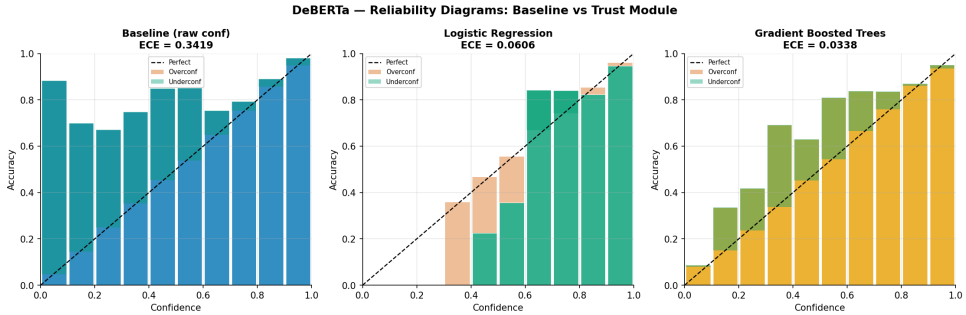

Figure 2: Reliability diagrams for DeBERTa: baseline raw confidence (left), Logistic Regression calibration corrector (centre), and Gradient Boosted Trees calibration corrector (right). ECE falls from 0.342 to 0.034 with GBT, and the reliability diagram moves from a flat, compressed pattern to one tracking the perfect calibration diagonal.

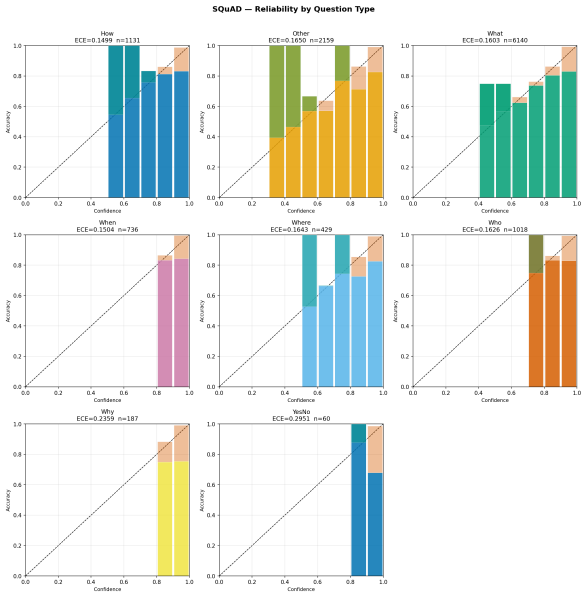

Figure 3: Reliability diagrams by question type for RoBERTa-base on SQuAD 2.0. In each panel, bars show accuracy per confidence bin (coloured), with orange shading indicating overconfidence (confi- dence exceeds accuracy) and the dashed diagonal rep- resenting perfect calibration. ECE ranges from 0.150 (How) to 0.295 (Yes/No), demonstrating that miscalibration is structured and predictable rather than random noise.

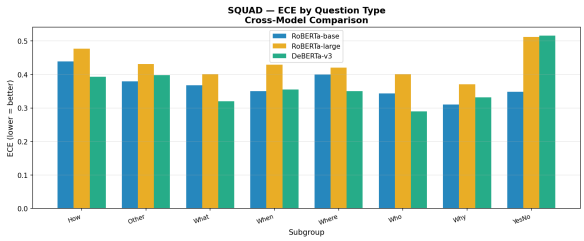

Figure 4: ECE by question type across all three extractive models. The subgroup ranking is consistent across architectures: Yes/No is the worst-calibrated type for all models. RoBERTa-large has the highest Yes/No ECE (0.512) while DeBERTa leads on other question types.

Model	Answerable		Unanswerable	
	ECE	Acc.(%)	ECE	Acc.(%)
RoBERTa-base	0.446	83.3	0.302	82.1
RoBERTa-large	0.517	87.0	0.323	89.6
DeBERTa-v3-base	0.670	86.6	**0.067**	88.0
Llama-3.2-3B	**0.080**	82.3	***0.615***	25.2

Table 5: ECE and accuracy split by answerability. **Bold** indicates best ECE per answerability condition. ***Bold italic*** highlights the worst ECE, marking the LLM abstention failure. Extractive models and the LLM show opposite failure modes.

on this subgroup. Of 1,472 unanswerable examples in the stratified sample, the model correctly abstained on only 371 (25.2%), generating a wrong answer on the remaining 1,101 (74.8%).

Extractive models show the exact opposite pattern, where abstention is handled reliably but answerable calibration is worse. These two failure modes are structurally complementary, suggesting that the two paradigms are miscalibrated in fundamentally different ways.

The verbalized confidence signal provides a striking additional finding. Table 6 shows that when the LLM correctly abstained, its mean verbalized confidence was 0.010, nearly zero. When it hallucinated an answer on an unanswerable question, mean verbalized confidence was 0.682, a gap of 0.672. By contrast, the same signal separates correct from incorrect answers on answerable questions by only 0.241. Verbalized confidence therefore functions primarily as a binary abstention detector rather than a graded confidence signal. The logprob confidence signal shows no such separation, assigning high token probability to both correct answers and hallucinated ones alike. These two signals are complementary rather than redundant, motivating their joint use in the calibration

Confident Hallucination on an Unanswerable Question

Question: What pope opposed Robert's campaign?
Context: *"Having obtained the consent of pope Gregory VII and acting as his vassal, Robert continued his campaign..."*
Predicted: Gregory VII **Ground truth:** unanswerable
Logprob confidence: 0.87 **Verbalized confidence:** 1.0

Gregory VII consented to Robert's campaign rather than opposing it. The passage does not answer the question as asked, yet the model extracted a plausible name and reported near-maximum confidence on both signals.

Figure 5: A real prediction from Llama-3.2-3B-Instruct on a SQuAD 2.0 unanswerable question.

corrector.

Condition	n	%	Mean Verbal Conf.
Answerable (n = 1,527)			
Correct answer	1,348	88.3	0.903
Incorrect answer	177	11.6	0.662
Unanswerable (n = 1,472)			
Correct abstention	371	25.2	**0.010**
Hallucinated answer	1,101	74.8	**0.682**

Table 6: Verbalized confidence by correctness and answerability. **Bold** highlights the two values driving the abstention detection gap of 0.672, nearly three times the answerable gap of 0.241.

5.5 Cross-Model Calibration Corrector

Table 7 reports global results for the calibration corrector evaluated on DeBERTa, which was held out entirely during training. GBT reduces DeBERTa's ECE from 0.342 to 0.034, a 90% reduction on a held-out architecture. LR also substantially improves ECE (0.342 to 0.061), though by a smaller margin. For AUC-RC, GBT achieves a modest improvement (0.068 to 0.064) while LR degrades it slightly (0.068 to 0.109). This asymmetry reflects an important distinction: ECE measures whether confidence values are numerically honest, while AUC-RC measures whether they rank predictions correctly for selective abstention. DeBERTa's raw confidence was already a reasonable ranker despite being poorly calibrated numerically, and LR's recalibration disrupts this ranking property while fixing the numerical calibration.

The subgroup breakdown confirms this generalisation holds across all question types and answer lengths. Answerable questions improve from ECE = 0.670 to 0.040 with GBT, and Yes/No questions improve from ECE = 0.516 to 0.200, specifically fixing the hard subgroups identified in Section 5.2. Calibration patterns learned on RoBERTa reflect general properties of extractive QA that transfer across architectures.

Method	ECE	ΔECE	AUC-RC	ΔAUC-RC
Baseline (raw conf.)	0.342	—	0.068	—
Logistic Regression	0.061	-0.281	0.109	+0.041
Gradient Boosted	**0.034**	**-0.308**	**0.064**	**-0.004**

Table 7: Calibration Corrector results on DeBERTa (never seen during training). **Bold** indicates the best performing method per metric. GBT reduces ECE by 90% while marginally improving AUC-RC. Negative Δ values indicate improvement for both metrics.

6 Conclusion

This paper presents Calibrated-QA-Bench, a systematic cross-model calibration audit for extractive question answering. Our central finding is that miscalibration in extractive QA is not random but structured: Yes/No questions are nearly twice as miscalibrated as How questions across all three model families tested, and this subgroup ranking is strikingly consistent regardless of model size or architecture. We further show that architectural choices produce predictable calibration fingerprints: DeBERTa is dramatically better calibrated on abstention decisions but dramatically worse on answerable questions compared to RoBERTa variants, a tradeoff with direct implications for model selection when one failure mode is costlier than the other.

The most striking finding emerges from the LLM comparison. Extractive models and generative LLMs are miscalibrated in precisely opposite ways: the LLM achieves the best answerable calibration of any model studied yet fails to abstain on 75% of unanswerable questions, while extractive models show the reverse — a structural difference in where each paradigm places its uncertainty, not simply a matter of one being better than the other. We further show that verbalized confidence from the LLM nearly perfectly separates correct abstentions from hallucinated answers, with a gap of 0.672 compared to only 0.241 for answer correctness on answerable questions. This asymmetry reveals that verbalized and logprob confidence are complementary signals measuring fundamentally different aspects of uncertainty, a finding with direct practical implications for how LLM confidence should be elicited and interpreted. Finally, a lightweight calibration corrector trained on RoBERTa predictions reduces DeBERTa's miscalibration by 90% on a held-out architecture, demonstrating that subgroup-aware calibration patterns transfer across model families without any retraining.

7 Limitations

The calibration corrector was evaluated on a single held-out architecture and should be tested across a broader set of model families to strengthen the generalisation claim. The LLM comparison used a small zero-shot baseline; a natural extension is to study instruction-tuned or fine-tuned generative models, where the abstention failure mode may be less severe. Finally, combining logprob and verbalized confidence signals in a joint calibration corrector is a promising direction that our findings directly motivate but that we leave for future work.

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696. Association for Computational Linguistics.

Laura Manor and Junyi Jessy Li. 2019. Plain english summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734. Association for Computational Linguistics.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2006. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595. Association for Computational Linguistics.