Cost-Sensitive Policy Selection for Large Language Models: A Failure Analysis of Post-Hoc Uncertainty Transformations

Ravi Teja Reddy Mandala
Oracle America Inc.
Austin, TX, USA
ORCID: 0009-0003-0313-4864

Bhavya Lahari Bommireddy
Microsoft Corporation
Redmond, WA, USA
ORCID: 0009-0006-2800-7812

***Abstract*—Large language models are increasingly used in decision-support systems where incorrect predictions can incur asymmetric costs. Conventional calibration aligns confidence with predictive correctness but does not directly optimize actions such as answering, abstaining, deferring, or invoking a tool. This paper investigates Decision-Calibrated Uncertainty (DCU) as a post-hoc utility-aware policy-selection framework rather than a demonstration of probabilistic calibration. We examine decision-aware temperature scaling and a Utility-Aware Correctness Head based on logits, entropy, and Monte Carlo dropout statistics. Experiments on SST-2 and SQuAD 2.0 show lower in-distribution decision cost under the selected cost matrices. However, the calibration head effectively learns an always-answer policy rather than improving instance-level uncertainty discrimination, and this policy incurs high cost under a constructed SST-2 negation-based stress test. The temperature-scaling implementation converges to a boundary solution whose optimality could not be independently verified. The study documents three failure modes: an unverified optimizer outcome, collapse to a constant policy, and fragility under a controlled stress test. They show that favorable average decision cost does not, by itself, establish improved probability calibration or instance-level uncertainty quality.**

***Index Terms*—Large Language Models, Uncertainty Quantification, Confidence Calibration, Cost-Sensitive Decision Making, Bayesian Decision Theory, Selective Prediction, Temperature Scaling, Monte Carlo Dropout, Distribution Shift, Decision-Calibrated Uncertainty**

I. Introduction

Large language models (LLMs) perform strongly across natural language processing tasks, but their confidence estimates may not reliably indicate whether individual predictions are correct. In deployed systems, a prediction may be answered, rejected, forwarded to a human, or used to invoke an external tool. The value of uncertainty therefore depends on both predictive calibration and the costs of the actions selected from it.

Most post-hoc methods remain prediction-centered. Temperature scaling minimizes negative log-likelihood [1]; self-consistency uses agreement across outputs [8]; and Monte Carlo (MC) dropout uses stochastic forward passes [4]. These methods do not directly optimize asymmetric action costs, so calibrated probabilities may still induce poor operational decisions.

We investigate a framework termed *Decision-Calibrated Uncertainty* (DCU), which connects estimated correctness scores to Bayesian cost-sensitive action selection. Despite the name, the evaluated implementations are post-hoc utility-aware score transformations and policy-selection mechanisms; the experiments do not establish that their outputs satisfy standard probabilistic calibration criteria. Its intended actions are answer, abstain, defer, and query-tool. However, query-tool is never uniquely optimal under the evaluated cost matrices, and neither query-tool nor defer is selected empirically. Under the SST-2 matrix, query-tool ties answer and abstain at exactly $p = 0.75$ but is otherwise weakly dominated.

Two post-hoc implementations are evaluated. Decision-Aware Temperature Scaling (DTS) selects a scalar temperature using empirical decision cost. An Utility-Aware Correctness Head (UCH) maps logits, entropy, and MC-dropout statistics to correctness probability. Experiments use frozen Flan-T5-Base and Flan-T5-Large models [14], [15] on SST-2 [16], [17] and SQuAD 2.0 [18].

The principal findings are:

- UCH lowers in-distribution decision cost by learning an always-answer policy.
- The same policy incurs high cost under a negation-based SST-2 stress test.
- The reported DTS result is an unverified optimizer outcome on a discontinuous objective.
- Lower decision cost does not by itself establish improved uncertainty quality.

II. Related Work

Post-hoc calibration aligns confidence with predictive correctness. Temperature scaling is a widely used baseline [1], while LLM uncertainty methods include verbalized confidence, consistency-based estimates, and probability-based techniques [11]. Language models can sometimes evaluate their own outputs, although calibration errors remain [6]. Calibration tuning modifies model parameters to improve

TABLE I. SST-2
DECISION COSTS

Action	Correct	Incorrect
Answer	0.0	10.0
Abstain	2.5	2.5
Defer	4.0	1.5
Query-tool	1.0	7.0

TABLE II. SQUAD 2.0
DECISION COSTS

Action	Correct	Incorrect
Answer	0.0	10.0
Abstain	3.0	3.0
Defer	5.0	2.0
Query-tool	2.0	6.0

generated confidence [2], and prior work has examined the relation between calibration and alignment [7].

MC dropout uses stochastic inference [4]; self-consistency uses agreement across reasoning paths [8]; and softmax scores remain common baselines for misclassification and out-of-distribution detection [9]. Selective prediction permits abstention when confidence is insufficient. Chow established the classification-rejection trade-off [13], and Geifman and El-Yaniv formalized risk–coverage analysis [3]. Selective question answering can degrade under domain shift [10], while self-evaluation has been studied for selective generation [12]. The fragility examined in this study is consistent with the observation of Kamath *et al.* [10] that confidence-based selection mechanisms trained in-domain can deteriorate under domain shift. Our negation-based construction is a controlled stress test, however, and is not equivalent to natural domain transfer. SelectLLM jointly considers answer quality and abstention during model adaptation [5]; DCU instead keeps the underlying Flan-T5 model frozen [14], [15] and evaluates SST-2 [16], [17] and SQuAD 2.0 [18].

Bayesian decision theory selects the action with minimum posterior expected loss. Applying it to LLM outputs makes asymmetric costs explicit, but lower realized cost does not necessarily imply improved probability calibration.

III. DECISION-THEORETIC FRAMEWORK

A. Problem Formulation and Action Space

Let a frozen model produce prediction $\hat{y}$ and confidence information for input x. The intended action space is

$$\mathcal{A} = \{\text{answer}, \text{abstain}, \text{defer}, \text{query-tool}\}. \quad (1)$$

Each action incurs cost $C(a \mid o)$, where $o \in \{\text{correct}, \text{incorrect}\}$. Tables I and II report the evaluated costs, which are illustrative assumptions rather than estimates from deployed applications.

Abstention means withholding an answer without routing the input elsewhere. Deferral means assigning the input to another human, model, or service. Because no downstream expert is modeled in this work, defer is represented only by a fixed illustrative cost. The costs are illustrative experimental assumptions rather than measurements from deployed systems and should not be interpreted as operational estimates.

The objective is to study how estimated correctness scores induce actions under asymmetric costs. This defines a cost-sensitive policy-selection problem that is related to, but distinct from, probability calibration. A Bayes-optimal action rule relies on accurate posterior correctness probabilities, whereas minimizing empirical decision cost alone does not establish that the resulting scores are calibrated.

B. Expected Decision Cost

Let $p = P(\text{correct} \mid x)$. The expected cost of action a is

$$R(a \mid p) = pC(a \mid \text{correct}) + (1-p)C(a \mid \text{incorrect}), \quad (2)$$

and the selected action is

$$a^*(p) = \arg\min_{a \in \mathcal{A}} R(a \mid p). \quad (3)$$

For SST-2, the decision rule selects defer for $p < 0.40$, abstain for $0.40 \leq p < 0.75$, and answer for $p \geq 0.75$. The corresponding SQuAD 2.0 thresholds are 0.334 and 0.70. For binary maximum-softmax confidence, the maximum class probability is necessarily at least 0.5. Therefore, the defer regions are structurally unreachable for greedy maximum-softmax confidence, NLL temperature scaling, DTS, and MC mean-confidence when the mean is computed from binary maximum-softmax scores. The UCH output is not subject to this lower bound, but it also never enters the defer region empirically. Query-tool is never uniquely optimal under either evaluated cost matrix. Under the SST-2 matrix, it ties answer and abstain at $p = 0.75$ but is otherwise weakly dominated. Consequently, the observed evaluation reduces primarily to answer-versus-abstain selective prediction and does not empirically test effective selection among four competitive actions.

C. Decision-Aware Temperature Scaling

Given logits $\mathbf{z}_i$, temperature-scaled confidence is

$$\hat{p}_i(T) = \max_k \left[\text{softmax}\left(\frac{\mathbf{z}_i}{T}\right)\right]_k. \quad (4)$$

DTS minimizes empirical realized decision cost:

$$T^*_{\text{DTS}} = \arg\min_{T \in [0.1, 20]} \frac{1}{N} \sum_{i=1}^{N} L_{\text{real}}\left(a^*(\hat{p}_i(T)), o_i\right), \quad (5)$$

where o_i indicates whether prediction i is correct. Because actions change only at confidence thresholds, the objective is piecewise constant and non-smooth.

D. Reproducibility Details

All experiments use random seed 42. The frozen backbone models are Flan-T5-Large and Flan-T5-Base. UCH uses five features: the two candidate-label log probabilities, MC-dropout mean confidence, MC-dropout confidence standard deviation, and predictive entropy. The hidden-layer width is 32, dropout rate is 0.1, MC-dropout uses 8 stochastic forward passes, and optimization uses Adam with learning rate 10^{-3} for 400 epochs.

E. Utility-Aware Correctness Head

The Utility-Aware Correctness Head (UCH) is a deterministic two-layer multilayer perceptron that maps post-hoc uncertainty features to a correctness score for cost-sensitive action selection. It is not a Bayesian neural network, and its output is not assumed to be probabilistically calibrated. It receives

$$\mathbf{u} = [\ell_0, \ell_1, c_{\mathrm{MC}}, \sigma_{\mathrm{MC}}, H], \tag{6}$$

where ℓ_0 and ℓ_1 are class logits, c_{MC} and σ_{MC} are the mean and standard deviation of MC-dropout confidence, and H is predictive entropy. For each action, the expected decision cost is computed from the predicted correctness score $\hat{p}$. A differentiable approximation of minimum-risk action selection is implemented using SoftMin weighting:

$$w_a(\hat{p}) = \frac{\exp(-R(a|\hat{p})/\tau)}{\sum_{b\in\mathcal{A}} \exp(-R(b|\hat{p})/\tau)} \tag{7}$$

$$\widetilde{R}(\hat{p}) = \sum_{a\in\mathcal{A}} w_a(\hat{p}) R(a|\hat{p}) \tag{8}$$

The training objective is

$$\mathcal{L}_{\mathrm{UCH}} = \widetilde{R}(\hat{p}) + \lambda\,\mathrm{BCE}(\hat{p}, y_c) \tag{9}$$

where $y_c = 1$ when the frozen model's prediction is correct and $y_c = 0$ otherwise.

We use $\tau = 0.1$, $\lambda = 1$, and a two-layer MLP with an input dimension of five, a hidden dimension of 32, ReLU activations, and a sigmoid output layer. The network includes a dropout layer with rate 0.1 between the first and second hidden layers and is trained using Adam for 400 epochs with a learning rate of 10^{-3}. MC-dropout uncertainty is estimated from 8 stochastic forward passes with dropout enabled during inference. Because the cost term depends on the head's own estimate, it may reward a globally favorable action region without improving instance-level ranking. Unlike the standard SoftMin operator, which returns a vector of normalized weights, Eq. (8) defines a scalar differentiable surrogate risk obtained from SoftMin-weighted expected cost.

Because we do not evaluate a BCE-only head with the same inputs, class-balanced BCE, alternative values of λ or τ, action-entropy regularization, or a minimum-abstention constraint, the present experiment cannot isolate whether policy collapse is caused by the decision-cost term, calibration-set composition, optimization, or limited information in the uncertainty features.

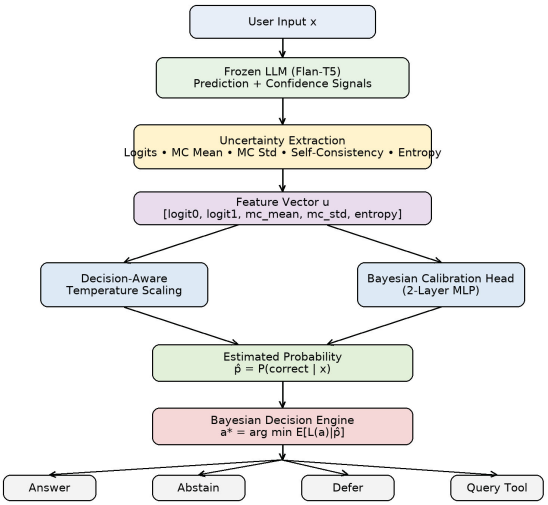


Fig. 1. Cost-sensitive policy-selection framework. The Flan-T5 backbone remains frozen while only the Utility-Aware Correctness Head is trained.

F. Decision Pipeline

Figure 1 summarizes the framework. A frozen LLM produces predictions and uncertainty signals. DTS or UCH transforms these signals into an estimated correctness probability, after which Eq. (3) selects the minimum expected-cost action. The transformations are post hoc and leave the base model unchanged.

IV. Experimental Setup

SST-2 is treated as binary sentiment classification, and SQuAD 2.0 is reduced to binary answerability detection. SST-2 calibration examples are sampled from the GLUE SST-2 training split using random seed 42. In-distribution evaluation uses the full SST-2 validation set (872 examples). The SST-2 stress test applies a deterministic negation transformation to every validation example and flips the corresponding sentiment label. The calibration set contains 200 examples and the evaluation set contains 872 examples. For SQuAD 2.0, an example is labeled answerable when the dataset contains at least one reference answer and unanswerable otherwise. The calibration set contains 150 examples sampled from the training split using random seed 42. The in-distribution evaluation set contains 1000 examples sampled from the validation split.

We evaluate greedy softmax confidence, likelihood-based temperature scaling, self-consistency, MC-dropout mean confidence, DTS, and UCH. An oracle is included only as a reference lower bound. Actions are selected through Eq. (3). Always abstain costs 2.5 on SST-2 and 3.0 on SQuAD 2.0. Always-answer cost is

$$\bar{C}_{\mathrm{answer}} = (1 - \mathrm{accuracy})C(\mathrm{answer} \mid \mathrm{incorrect}). \tag{10}$$

The evaluation covers two binary tasks and two sizes from one encoder–decoder model family. It does not cover decoder-only models, multiclass prediction, open-ended generation, or a setting where all four actions are competitive.

A. Prompt Templates and Label Mapping

For SST-2, the prompt asks whether a sentence is positive or negative, using candidate labels "positive" and "negative." For SQuAD 2.0 answerability detection, the prompt asks whether a question can be answered from the passage, using labels "yes" and "no." Candidate-label sequence losses are normalized into two-class probabilities, and confidence is defined as the maximum normalized probability. MC-dropout uncertainty is estimated from 8 stochastic forward passes.

V. Results

A. Base Model and Uncertainty Statistics

Flan-T5-Large obtains calibration and evaluation accuracies of 0.930 and 0.943 on SST-2, with mean confidence approximately 0.537. Under the constructed SST-2 stress test, accuracy decreases to 0.078 while confidence changes little. For SQuAD 2.0, calibration and evaluation accuracies are 0.840 and 0.806, respectively, with mean confidence approximately 0.523. The mean MC-confidence standard deviation is 0.0183 on SST-2 and 0.0154 on SQuAD 2.0. Mean self-consistency is 0.847 ± 0.154 and 0.834 ± 0.155, respectively. These statistics show feature variation but do not establish useful error discrimination.

Under the constructed SST-2 negation-based shift, accuracy decreases from 0.913 to 0.073 while confidence changes little. Because the shift is synthetically generated, interpretation of this result depends on the validity of the corresponding label transformations. The SQuAD shift is not included in the central cost comparison.

B. Cost Sensitivity Analysis

We performed a secondary sensitivity study by varying the incorrect-answer penalty from 5 to 50. Across all settings, the relative ordering of methods remained qualitatively similar, although absolute costs changed substantially. This confirms that policy behavior depends directly on the chosen cost matrix.

C. Expected Decision Cost

Table III reports average decision costs across all evaluated methods. The UCH obtains an average cost of 0.867 on SST-2 ID and 1.500 on SQuAD ID. Relative to the always-abstain policy, these values correspond to numerical reductions of 65.3% and 50.0%, respectively. These percentages should not be interpreted as evidence of improved instance-level calibration. On SST-2, an always-answer policy has an approximate cost of $(1 - 0.913) \times 10 = 0.870$, which is effectively the same as the reported UCH cost of 0.867 after accounting for rounding of the reported accuracy. On SQuAD, the always-answer cost is $(1 - 0.850) \times 10 = 1.500$, exactly matching the reported UCH cost. The action distributions confirm that UCH answers every evaluated in-distribution example. Its apparent in-distribution gain therefore comes from selecting a globally favorable constant action under the chosen cost matrix, not from using confidence to distinguish safer examples from riskier examples. Under the SST-2 distribution shift, the same behavior becomes harmful. The always-answer cost is approximately $(1 - 0.073) \times 10 = 9.270$, which is again effectively identical to the reported UCH cost of 9.267. This comparison shows that UCH tracks the risk of an always-answer policy rather than adapting its actions in response to the shift. Action frequencies showed that UCH answered all evaluated in-distribution examples, whereas greedy confidence, MC confidence, and the reported DTS output abstained on all examples.

D. AUROC for Error Detection

Beyond decision cost, we evaluate the capacity of uncertainty estimates to distinguish correct predictions from errors.

Because AUROC orientation was not independently revalidated after the original experiment, no substantive conclusion depends on the exact value 0.166. The result is consistent with either poor correctness ranking or a score-orientation issue and should therefore be interpreted cautiously. Together with its all-answer action distribution, this result shows that the lower UCH decision cost arises from policy selection rather than improved uncertainty estimation. Although UCH attains an OOD AUROC of 0.744, its corresponding OOD decision cost remains high, demonstrating that ranking quality and decision quality are distinct properties. Confidence is computed from normalized sequence-level probabilities over competing label strings rather than from a dedicated classifier head. Consequently, high classification accuracy can coexist with confidence values close to 0.5. Confidence intervals for average decision cost were not computed, and the original Wilcoxon comparison primarily contrasts an all-answer policy with an all-abstain policy.

E. Distribution-Shift Robustness

Distribution-shift behavior is summarized in Table V. UCH increases from 0.867 to 9.267 under the SST-2 shift, effectively matching the always-answer change from 0.870 to 9.270. Its in-distribution gain and out-of-distribution failure are therefore consequences of the same aggressive policy. Methods that abstain on all examples retain constant cost because they avoid prediction risk, not because they detect the shift.

F. Cross-Model Comparison

Table VI compares results across Flan-T5-Large and Flan-T5-Base. Both models are evaluated on SST-2 using identical calibration procedures. Flan-T5-Large achieves an evaluation accuracy of 0.913, whereas Flan-T5-Base achieves lower accuracy under the same protocol. Although UCH reduces SST-2 cost for both model sizes, the gains diminish on the smaller Flan-T5-Base backbone. On SQuAD, UCH performs worse than both the Flan-T5-Large result and the always-abstain baseline. The inconsistent behavior across only two model sizes indicates that broader architectural generalization cannot be established from the current evaluation.

TABLE III. AVERAGE REALIZED DECISION COST ON THE REPORTED TEST SETS

Method or policy	SST-2 ID	SST-2 OOD	SQuAD ID
Always abstain	2.500	2.500	3.000
Always answer	0.870	9.270	1.500
Greedy confidence policy	2.500	2.500	3.000
NLL temperature scaling	0.867	7.833	1.810
Self-consistency policy	0.908	8.867	1.600
MC mean-confidence policy	2.500	2.500	3.000
DTS optimizer output	2.500	2.500	3.000
Utility-Aware Correctness Head	0.867	9.267	1.500
Oracle	0.130	1.390	0.300

TABLE IV. AUROC FOR CORRECTNESS DISCRIMINATION

Method	SST-2 ID	SST-2 OOD	SQuAD ID
Greedy confidence	0.888	0.117	0.710
NLL temperature scaling	0.888	0.117	0.710
Self-consistency	0.768	0.157	0.598
MC mean confidence	0.802	0.082	0.611
DTS optimizer output	0.888	0.117	0.710
UCH	0.166	0.744	0.517

TABLE V. DECISION-COST DEGRADATION UNDER DISTRIBUTION SHIFT

Method	ID Cost	OOD Cost	Δ
Greedy confidence	2.500	2.500	+0.000
NLL temperature scaling	0.867	7.833	+6.967
Self-consistency	0.908	8.867	+7.958
DTS optimizer output	2.500	2.500	+0.000
UCH	0.867	9.267	+8.400
Oracle	0.130	1.390	+1.260

TABLE VI. CROSS-MODEL COMPARISON

Model	Method	SST-2	SQuAD
Flan-T5-Large	Greedy	2.500	3.000
Flan-T5-Large	UCH	0.867	1.500
Flan-T5-Base	Greedy	2.500	3.000
Flan-T5-Base	UCH	1.000	3.067

VI. DISCUSSION

A. Decision cost versus uncertainty quality

The central empirical finding is not that UCH improves uncertainty calibration. Its SST-2 and SQuAD ID costs are equal, to the reported precision, to the corresponding always-answer costs. Its action distribution confirms that it answers every ID example, and its SST-2 ID AUROC of 0.166 indicates poor instance-level ordering of correctness. The observed cost reductions relative to greedy confidence therefore arise primarily because the selected cost matrices favor answering when average model accuracy is high. This distinction is important because policy selection and uncertainty estimation address different questions. A globally favorable constant action may minimize average cost without identifying which individual predictions are likely to be incorrect. Evaluations of decision-aware uncertainty should therefore report constant-action baselines together with probability-calibration metrics and selective-prediction curves.

B. Unverified optimization outcome for scalar DTS

The DTS implementation returned $T = 20$ and universal abstention; however, the optimization result was not independently verified through exhaustive search. Because actions change only when confidence crosses decision thresholds, the objective is piecewise constant. Verifying the optimum would require exhaustive grid search or breakpoint enumeration on the calibration-set objective. Consequently, the reported result should be interpreted as an unverified optimizer output rather than evidence against decision-aware temperature scaling.

C. Restricted action space

The formal framework includes answer, abstain, defer, and query-tool actions, but the empirical evaluation does not exercise all four. Query-tool is never uniquely optimal under either cost matrix. Under the SST-2 matrix, it ties answer and abstain at exactly $p = 0.75$ but is otherwise weakly dominated; under the SQuAD 2.0 matrix, another action has lower or equal expected cost throughout the relevant probability range. Deferment is analytically available at low correctness probabilities but is never selected by the evaluated methods. The experiment therefore reduces to answer-versus-abstain selective prediction. Demonstrating the broader action-space motivation would require cost structures and auxiliary systems for which deferment and tool invocation have empirically distinct outcomes. For example, tool use would require measured tool accuracy and latency or monetary cost, while deferment would require measured human accuracy, turnaround time, or workload. Those experiments are outside the scope of the present study.

D. Distribution-shift failure

The constructed SST-2 negation-based shift reduces measured model accuracy from 0.913 to 0.073 while leaving the uncertainty-feature distribution largely unchanged. Since this shift is synthetically generated, the result should be viewed

as an exploratory stress test rather than a characterization of natural domain shift. UCH consequently continues to answer almost every example. Its ID advantage and OOD failure are thus two consequences of the same learned policy rather than independent phenomena. This result shows that a post-hoc head trained only on ID examples cannot be assumed to recognize a change in the relationship between its features and correctness. Explicit shift detection, OOD calibration data, or conservative policy constraints would be required before considering such a method for changing deployment environments.

VII. Limitations

The results should be interpreted as exploratory. First, the DTS optimum was not verified through exhaustive search. Second, lower decision cost does not imply improved calibration because ECE, Brier score, reliability diagrams, NLL, and risk–coverage curves were not evaluated. Third, confidence intervals for decision costs were not computed and calibration sets were small. Fourth, only answer and abstain were used empirically because defer was unreachable and query-tool never became uniquely optimal under the selected costs. Finally, the evaluation covers only SST-2, binary SQuAD answerability, and two sizes of Flan-T5; broader generalization remains untested.

VIII. Conclusion

This paper evaluated a post-hoc utility-aware policy-selection framework for connecting language-model confidence estimates to cost-sensitive actions. The findings support interpreting the evaluated methods as policy-selection mechanisms rather than uncertainty-calibration methods in the conventional probabilistic sense. The framework emphasizes an important distinction: a confidence transformation should be evaluated not only by predictive criteria but also by the actions and costs it induces. The results highlight an important distinction between decision quality and uncertainty quality. UCH reduces decision cost relative to always abstain under the selected cost matrices, but its behavior is effectively equivalent to always answering and does not correspond to improved instance-level uncertainty estimation. The same aggressive policy incurs high decision cost under the constructed negation-based stress test. The temperature-scaling experiment also reveals a methodological problem. The DTS experiment produced a boundary solution whose calibration-set optimality was not independently verified. The result should therefore be interpreted as an unvalidated optimization outcome rather than evidence against decision-aware scalar temperature scaling in general. The main contribution of the study is consequently an analysis of failure modes in post-hoc cost-sensitive uncertainty processing. Future work should use global optimization for discontinuous action objectives, include constant-action baselines and confidence intervals, report probability-calibration and risk–coverage measures, evaluate multiple cost matrices, and construct settings in which deferment and tool use are meaningfully competitive. Broader evaluation across decoder-only models, multiclass tasks, and open-ended generation is also required before general claims can be made.

References

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1321–1330.

[2] S. Kapoor, N. Gruver, M. Roberts, A. Pal, S. Dooley, M. Goldblum, and A. G. Wilson, "Calibration-tuning: Teaching large language models to know what they don't know," in *Proc. 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP)*, 2024.

[3] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, 2016, pp. 1050–1059.

[5] Y. Mao, T. Durand, N. Mehrasa, J. He, and M. Ester, "SelectLLM: Calibrating LLMs for selective prediction: Balancing coverage and risk," unpublished manuscript, 2026.

[6] S. Kadavath *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[7] C. Zhu, B. Xu, Q. Wang, and Y. Zhang, "On the calibration of large language models and alignment," in *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

[8] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain-of-thought reasoning in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[9] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[10] A. Kamath, R. Jia, and P. Liang, "Selective question answering under domain shift," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 5684–5696.

[11] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A survey of confidence estimation and calibration in large language models," *arXiv preprint arXiv:2311.08298*, 2024.

[12] J. Ren, Y. Zhao, T. Vu, P. J. Liu, and B. Lakshminarayanan, "Self-evaluation improves selective generation in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[13] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.

[14] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[15] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[16] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of EMNLP*, 2013, pp. 1631–1642.

[17] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[18] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 784–789.