Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Hierarchy-Aware Supervised Uncertainty Estimation for Black-box LLM Taxonomic Reasoning

Shuting Xie [†], Nathaniel Lesperance [‡, †], Graham W. Taylor [‡,†*]
[†] Vector Institute for AI, Toronto, ON
[‡] University of Guelph, Guelph, ON

Abstract

Large language models (LLMs) are increasingly used for scientific decision support, yet reliable confidence estimation remains difficult in black-box settings. We study uncertainty estimation for hierarchical taxonomic reasoning generated by a black-box LLM in a long-tailed biodiversity monitoring pipeline. Using proxy features extracted by an open-source tool LLM, we train lightweight supervised estimators with hierarchy-aware supervision to predict rank-wise correctness. Across three tool LLMs, the supervised estimators consistently outperform a token-likelihood baseline for micro discrimination and selective prediction under a single global rejection threshold, improving micro AUROC from 0.57 to 0.75–0.80. The best results are achieved by a rank-specific multi-head design (H3), suggesting that accounting for hierarchical output structure is important when a unified abstention rule is required.

Keywords: Uncertainty, Large Language Model, Hierarchical Classification

1. Introduction

Large language models (LLMs) are increasingly deployed for reasoning and decision support in high-stakes domains such as medicine, finance, and ecology [1–3], where reliable uncertainty quantification (UQ) is as important as predictive accuracy. Yet estimating confidence for black-box LLMs, accessible only through APIs, remains an open problem. The problem is further complicated in hierarchical reasoning, where outputs span multiple semantic levels and abstention can yield partial outputs rather than a single flat label.

Biodiversity monitoring provides a particularly challenging instance of this setting. Taxonomic predictions are hierarchical and strongly long-tailed [4]. Rare and endangered species are difficult to monitor and can play disproportionately important roles in ecosystems [5], making UQ essential for routing uncertain cases to domain experts. A recent system [6] combines vision language model (VLM)-based image captioning, retrieval-augmented generation (RAG), and LLM-based reasoning to generate interpretable predictions across taxonomic ranks. However, it lacks numerical confidence estimates, relying instead on qualitative abstentions that do not support principled rejection thresholds.

We address this gap by studying uncertainty estimation for black-box LLMs in hierarchical taxonomic reasoning, where only externally accessible signals are available. Our contributions are threefold:

(1) **Formulation.** We cast black-box LLM taxonomic UQ as rank-level selective prediction under partial-path abstention.
(2) **Method.** We develop lightweight supervised uncertainty estimators that use proxy-LLM features, hierarchy-aware supervision, and rank-aware output parameterizations to predict rank-wise correctness.
(3) **Design insight.** On a realistic long-tailed arthropod pipeline, supervised estimators outperform a token-likelihood baseline for global-threshold selective prediction, and a rank-specific multi-head design performs best overall.

* gwtaylor@uoguelph.ca

2. Related Work

Uncertainty estimation (UE) for LLMs. LLM uncertainty estimation has been studied in both white-box and black-box settings. White-box methods leverage logits or hidden states [7], but these signals are unavailable for proprietary API models. Black-box approaches typically rely on self-verbalized confidence [8] or consistency or semantic clustering across multiple sampled responses [9, 10]. The former may fail to faithfully reflect the model's intrinsic uncertainty [11], while the latter can be prohibitively expensive in API-based pipelines. We therefore build on the supervised black-box approach of Liu *et al.* [12], which uses an open-source *tool LLM* to extract proxy features from prompt-response pairs and trains a correctness predictor for the target model. We view tool-LLM features as a practical proxy signal rather than faithful reconstruct of target uncertainty.

Hierarchical reasoning in taxonomic prediction. Taxonomic reasoning is a challenging setting for uncertainty estimation because predictions are hierarchical, long-tailed, and often partially specified through abstention. Prior work on hierarchical classification has developed evaluation measures based on ancestor overlap and partial credit [13], and recent work has extended this perspective to taxonomy-aware evaluation [14]. In parallel, an LLM-based pipeline has been proposed for taxonomic reasoning in rare arthropods [6]. Our focus is complementary. Rather than improving the upstream reasoning pipeline, we study UQ for its hierarchical outputs and use hierarchy-aware signals as auxiliary supervision.

3. Methodology

We formulate uncertainty estimation as a supervised prediction problem (Figure 1).

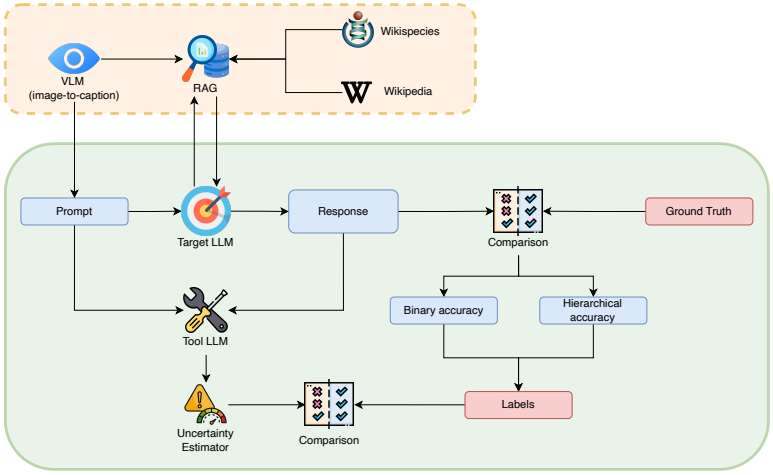


Figure 1. End-to-end pipeline for LLM uncertainty estimation. A target LLM generates taxonomic predictions from VLM captions and RAG-retrieved context. A tool LLM extracts proxy features from each prompt–response pair, which a lightweight uncertainty estimator maps to a confidence score. Blue boxes denote inputs and intermediate outputs. Red boxes denote prediction targets.

3.1. Dataset and Upstream Taxonomic Reasoning System

This study uses 829 arthropod images from the Rare Species dataset [15], obtained after filtering invalid samples from an initial set of 951. Each sample is associated with a complete

taxonomic path, spanning seven taxonomic ranks. The dataset is conservation-biased, with all species belonging to International Union for Conservation of Nature (IUCN) Red List categories from Near Threatened to Extinct in the Wild.

As input to our uncertainty estimator, we use the textual taxonomic outputs from the upstream simple-RAG reasoning system [6], instantiated with GPT-4o [16] as the target LLM. The system predicts taxa in a coarse-to-fine manner and abstains at the first uncertain rank, producing variable-depth partial paths and consequently limited coverage at finer ranks (Appendix Table 2). We therefore model uncertainty at the rank level to capture both error propagation and partial-path abstention.

3.2. **Supervision Labels**

Each rank-level instance has two supervision signals:(i) *a binary accuracy* $y \in \{0, 1\}$, indicating whether the predicted taxon matches the ground truth at that rank; and (ii) *a hierarchical accuracy* $h \in [0, 1]$, which assigns partial credit based on taxonomic overlap. Following [14], we compute hierarchical precision, recall, and F1. For predicted node v^{pred} and gold node v^{gold}, letting $\text{anc}(v)$ denote the set of ancestors of v, including v itself,

$$hP = \frac{|\text{anc}(v^{\text{pred}}) \cap \text{anc}(v^{\text{gold}})|}{|\text{anc}(v^{\text{pred}})|}, \quad hR = \frac{|\text{anc}(v^{\text{pred}}) \cap \text{anc}(v^{\text{gold}})|}{|\text{anc}(v^{\text{gold}})|}, \quad hF = \frac{2hPhR}{hP + hR}. \tag{3.1}$$

We set $h = hF$ and use it only as an auxiliary training target. Kingdom and phylum are excluded, as they are constant across all samples and hence non-discriminative.

3.3. **Tool LLM Feature Extraction**

Following Liu *et al.* [12], we use an open-source tool LLM to extract proxy features from each prompt-response pair. We consider Gemma 7B [17], GPT-OSS 20B [18], or Qwen3 30B A3B Instruct [19] as tool LLMs (layer details in Appendix Table 3). We collect two feature families (Appendix Figure 3): (i) *representation features*, namely averaged and last-token hidden states from the middle layer of the tool LLM, and (ii) *distributional features*, including statistics of entropy, negative log-probability, and raw token probability.

3.4. **Supervised Uncertainty Estimators**

To mitigate overfitting under sparse fine-rank supervision, we use a compact MLP with a shared two-layer trunk (hidden size 16, ReLU, dropout 0.3) and task-specific head(s). The primary head is a sigmoid classifier that predicts the probability of rank-wise correctness, directly supporting AUROC and risk–coverage evaluation. Some variants add a regression head for hierarchical F1 to provide hierarchy-aware auxiliary supervision.

We train the primary head with binary cross-entropy (BCE) and the auxiliary head with mean squared error (MSE):

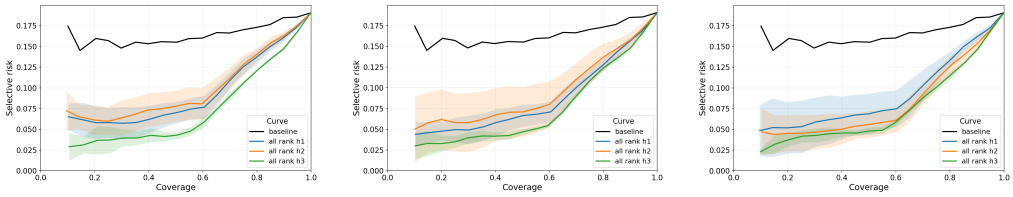
$$\mathcal{L} = \underbrace{\text{BCE}(\hat{p}, y)}_{\mathcal{L}_{\text{BCE}}} + \underbrace{\lambda \|\hat{h} - h\|_2^2}_{\mathcal{L}_{\text{MSE}}}. \tag{3.2}$$

Here, $\hat{p}$ denotes the predicted probability of rank-wise correctness, y the binary correctness label, $\hat{h}$ the predicted hierarchical score, and h the hierarchical-F1 auxiliary target.

We study three variants: **H1**, a shared MLP trunk with a single binary head trained with $\mathcal{L}_{\text{BCE}}$ only; **H2**, the same trunk augmented with an auxiliary regression head and trained with Eq. (3.2); and **H3**, a shared trunk with five rank-specific binary heads plus one auxiliary regression head. BCE is applied only to the head matching the sample rank, while the auxiliary regression loss is applied globally. At inference, H3 produces correctness probabilities for all five ranks in a single forward pass. Implementation details are provided in Appendix A.

Table 1. All-rank micro AUROC (mean ± std over 5 seeds). Highest mean is bolded.

Tool LLM	Benchmarks	Ours		
	Baseline	H1	H2	H3
Gemma 7B		0.710 ± 0.050	0.694 ± 0.060	**0.796 ± 0.023**
Qwen3 30B A3B Instruct	0.573	0.743 ± 0.050	0.719 ± 0.080	**0.749 ± 0.017**
GPT-OSS 20B		0.733 ± 0.069	0.783 ± 0.044	**0.790 ± 0.028**



(a) Gemma 7B. *(b)* Qwen3 30B A3B Instruct. *(c)* GPT-OSS 20B.

Figure 2. Global-threshold risk–coverage curve. Lower is better. Curves show the mean over 5 random seeds, with shaded regions indicating ±1 standard deviation.

3.5. **Negative log-likelihood (NLL) baseline**

As a training-free baseline, we use the negative mean token NLL of the target LLM output [20, 21]. For generated output $\mathbf{y} = (y_1, \ldots, y_T)$ under context $\mathbf{c}$, we define

$$\mathrm{conf}_{\mathrm{nll}}(\mathbf{y} \mid \mathbf{c}) = -\mathrm{NLL}_{\mathrm{mean}}(\mathbf{y} \mid \mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t \mid \mathbf{c}, y_{<t})\,, \tag{3.3}$$

where $p_\theta(y_t \mid \mathbf{c}, y_{<t})$ denotes the token probability returned by the target LLM, and $T = |\mathbf{y}|$ is the output length in tokens. We use this baseline because it is available in a single GPT-4o API pass, whereas full-distribution or multi-sample baselines are unavailable or costly.

3.6. **Evaluation protocol**

We evaluate uncertainty at the rank level. Each image–rank pair (x, r) is one example, with correctness label $y_r^{\mathrm{corr}}(x) = \mathbf{1}[\hat{y}_r^{\mathrm{tax}}(x) = y_r^{\mathrm{tax}}(x)]$ and confidence score $\mathrm{conf}(x, r)$.

Risk–coverage Curve. To reflect deployment under a single global rejection rule, we pool validation examples across ranks and, for a target coverage c, choose a threshold $\tau = \tau(c)$ as the $(1-c)$-quantile of confidence scores. On the test set, we accept examples with $\mathrm{conf}(x, r) \geq \tau$, and define selective risk as the error rate among accepted examples [22]:

$$R(\tau) = \frac{\mathbb{E}\big[\big(1 - y_r^{\mathrm{corr}}(x)\big)\,\mathbf{1}[\mathrm{conf}(x, r) \geq \tau]\big]}{\mathbb{E}[\mathbf{1}[\mathrm{conf}(x, r) \geq \tau]]}. \tag{3.4}$$

AUROC. We report micro AUROC (pooled across ranks) and macro AUROC (per-rank average) as threshold-free discrimination metrics, with $y_r^{\mathrm{corr}}(x) = 1$ as the positive class.

4. **Results**

Table 1 shows that supervised uncertainty estimators consistently and substantially outperform the mean-NLL baseline in micro AUROC across all three tool LLMs. H3 achieves the highest mean micro AUROC overall, suggesting the strongest discrimination between

correct and incorrect, although its margin over H1 is small for Qwen3 and clearer for Gemma 7B and GPT-OSS 20B.

Figure 2 provides the most deployment-relevant comparison under a single global rejection threshold. Across all three tool LLMs, our supervised uncertainty estimators achieve remarkably lower selective risk than the mean-NLL baseline over a broad coverage range, with the largest gains at low-to-mid coverage. H3 consistently yields the best overall risk–coverage trade-off, while H1 and H2 also remain better than the baseline throughout.

The gains are less consistent under rank-balanced macro AUROC metric (Appendix Table 4 and Figure 4). Calibration results are also reported in Appendix Table 5 and generally favour the supervised estimators over the mean-NLL baseline.

5. Discussion

The divergence between micro (Table 1) and macro AUROC (Appendix Table 4) stems from rank imbalance. Micro AUROC is dominated by high-coverage coarse ranks, whereas macro AUROC weights each rank equally. Per-rank results (Appendix Figure 4) suggest that the degradation is concentrated at deeper ranks, where frequent abstentions reduce the available training signal and increase variance (Appendix Table 2). Therefore, the rank-balanced metric exposes weaknesses that the deployment-oriented micro metric downweights.

For unified-threshold deployment, cross-rank comparability is key. H2 yields modest gains over H1 for most tool LLMs, except with GPT-OSS 20B, while H3 performs best overall (Figure 2). Since H2 and H3 use the same auxiliary supervision, this pattern suggests that rank-specific output heads matter more than the auxiliary hierarchical target alone.

A main limitation is sparse fine-rank supervision. Our study is further restricted to a single taxonomic reasoning pipeline and a global abstention threshold chosen to reflect deployment constraints. Future work should assess whether these conclusions generalize to richer fine-rank coverage, wider distribution shifts, and rank-specific decision rules.

6. Conclusion

We studied UQ for black-box LLMs in hierarchical taxonomic reasoning under long-tailed data and partial-path abstention. Using proxy features extracted by a tool LLM, we trained lightweight supervised estimators with hierarchy-aware signals to predict rank-wise correctness. Across tool LLMs, these estimators outperformed a NLL baseline in micro discrimination and global-threshold selective prediction. A rank-specific multi-head design performed best overall, suggesting that estimator design for hierarchical outputs is as important as the uncertainty signal itself in unified-threshold deployment.

References

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. “Large language models encode clinical knowledge”. In: *Nature* 620.7972 (2023), pp. 172–180.

[2] Y. Li, S. Wang, H. Ding, and H. Chen. “Large language models in finance: A survey”. In: *Proceedings of the fourth ACM international conference on AI in finance*. 2023, pp. 374–382.

[3] M. Mora-Cross and S. Calderon-Ramirez. “Uncertainty estimation in large language models to support biodiversity conservation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. 2024, pp. 368–378.

[4] Z. Gharaee, Z. Gong, N. Pellegrino, I. Zarubiieva, J. B. Haurum, S. Lowe, J. McKeown, C. Ho, J. McLeod, Y.-Y. Wei, et al. “A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 43593–43619.

[5] L. E. Dee, J. Cowles, F. Isbell, S. Pau, S. D. Gaines, and P. B. Reich. "When do ecosystem services depend on rare species?" In: *Trends in ecology & evolution* 34.8 (2019), pp. 746–758.

[6] N. Lesperance, S. Ratnasingham, and G. Taylor. "Taxonomic Reasoning for Rare Arthropods: Combining Dense Image Captioning and RAG for Interpretable Classification". In: *Proceedings of the Canadian Conference on Artificial Intelligence* (2025).

[7] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych. "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)*. Association for Computational Linguistics, June 2024, pp. 6577–6595.

[8] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. "Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 5433–5442.

[9] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. "Detecting hallucinations in large language models using semantic entropy". In: *Nature* 630.8017 (2024), pp. 625–630.

[10] P. Manakul, A. Liusie, and M. Gales. "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models". In: *Proceedings of the 2023 conference on empirical methods in natural language processing*. 2023, pp. 9004–9017.

[11] G. Yona, R. Aharoni, and M. Geva. "Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?" In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2024, pp. 7752–7764.

[12] L. Liu, Y. Pan, X. Li, and G. Chen. *Uncertainty estimation and quantification for LLMs: A simple supervised approach*. 2024. arXiv: 2404.15993.

[13] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. "Evaluation measures for hierarchical classification: a unified view and novel approaches". In: *Data Mining and Knowledge Discovery* 29.3 (2015), pp. 820–865.

[14] V. Snæbjarnarson, K. Du, N. Stoehr, S. Belongie, R. Cotterell, N. Lang, and S. Frank. "Taxonomy-aware evaluation of vision-language models". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 9109–9120.

[15] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, et al. "BioCLIP: A vision foundation model for the tree of life". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 19412–19424.

[16] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. *GPT-4o system card*. 2024. arXiv preprint: 2410.21276.

[17] Gemma Team. *Gemma: Open models based on gemini research and technology*. 2024. arXiv preprint: 2403.08295.

[18] Open AI. "gpt-oss-120b & gpt-oss-20b model card". In: (2025). arXiv preprint: 2508.10925.

[19] Qwen Team. *Qwen3 Technical Report*. 2025. arXiv: 2505.09388.

[20] Z. Lin, S. Trivedi, and J. Sun. "Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2024, pp. 10351–10368.

[21] N. Gupta, H. Narasimhan, W. Jitkrittum, A. S. Rawat, A. K. Menon, and S. Kumar. "Language Model Cascades: Token-Level Uncertainty And Beyond". In: *The Twelfth International Conference on Learning Representations*. 2024.

[22] Y. Geifman and R. El-Yaniv. "Selectivenet: A deep neural network with an integrated reject option". In: *International conference on machine learning*. PMLR. 2019, pp. 2151–2159.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

Code can be found at https://github.com/uoguelph-mlrg/hierarchy-aware-llm-uq.

Appendix A. Implementation details

All MLPs are trained with Adam (lr $= 10^{-3}$, weight decay $= 3 \times 10^{-4}$), batch size 128, for up to 200 epochs. We use early stopping on validation micro AUROC (patience $= 5$, min $\Delta = 10^{-4}$). For hierarchical supervision we set $\lambda = 0.1$. An ℓ_1 penalty (10^{-4}) is applied to the first linear layer weights (input-to-hidden) to encourage sparse use of input features. Feature selection is applied only to the hidden-state representation features, combining the top-100 features from each of Lasso regression, mutual information, and Pearson correlation with the target. We use 5-fold cross-validation. To avoid specimen leakage, all rank-level records derived from the same specimen are assigned to the same cross-validation fold.

Appendix B. Upstream taxonomic pipeline statistics

Table 2. Rank coverage and prediction accuracy in the upstream system before filtering.

Rank	Attempts	Coverage (%)	Accuracy	F1	HP	HR	HF
Class	950	100	0.966	0.973	0.981	0.981	0.981
Order	922	97	0.872	0.885	0.955	0.955	0.955
Family	745	78	0.693	0.710	0.914	0.914	0.914
Genus	223	23	0.453	0.532	0.865	0.865	0.865
Species	55	6	0.455	0.500	0.847	0.847	0.847

Appendix C. Tool LLM configurations and extracted features

Table 3. Details of tool LLMs. The first column reports the model name, followed by the exact checkpoint identifier in parentheses.

Tool LLM	Middle layer	Total Number of Layers	Hidden size
Gemma 7B (`google/gemma-7b`)	14	28	3072
Qwen3 30B A3B Instruct (`Qwen/Qwen3-30B-A3B-Instruct-2507-FP8`)	24	48	2048
GPT-OSS 20B (`openai/gpt-oss-20b`)	12	24	2880

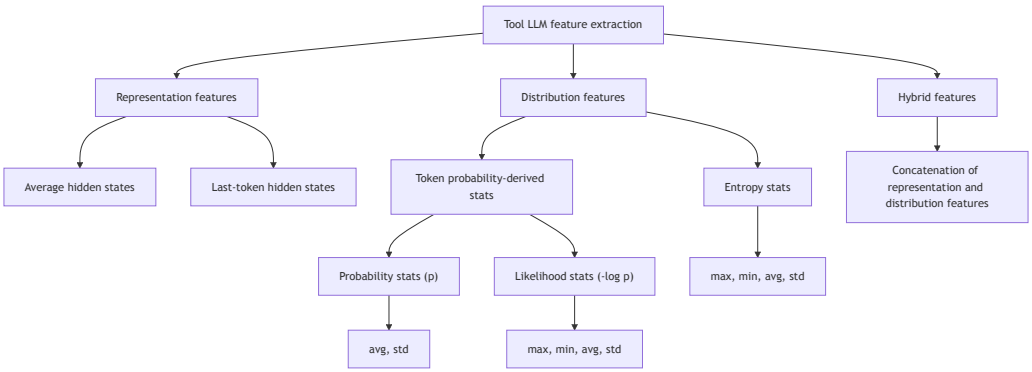
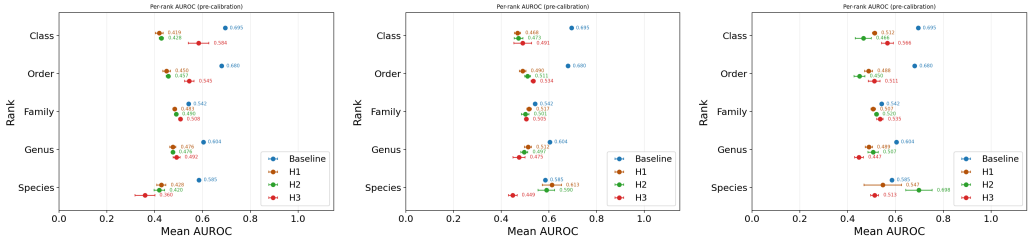


Figure 3. Feature tree demonstrates the feature types extracted by tool LLMs.

Appendix D. Additional UQ results

Table 4. All-rank aggregated macro AUROC across tool LLMs. Results are reported as mean ± std over 5 random seeds. Bold indicates the best macro AUROC across all methods, and underlining marks the best performance among our three uncertainty estimators.

Tool LLM	Benchmarks	Ours		
	Baseline	H1	H2	H3
Gemma 7B	**0.621**	0.451 ± 0.009	0.454 ± 0.013	<u>0.498 ± 0.036</u>
Qwen3 30B A3B Instruct		<u>0.520 ± 0.015</u>	0.514 ± 0.016	0.491 ± 0.021
GPT-OSS 20B		0.509 ± 0.029	<u>0.528 ± 0.044</u>	0.515 ± 0.031



(a) Gemma 7B. *(b)* Qwen3 30B A3B Instruct. *(c)* GPT-OSS 20B.

Figure 4. Per-rank AUROC (pre-calibration). Markers show the mean over 5 random seeds, and horizontal bars indicate ±1 standard deviation.

Table 5. All-rank calibration metrics before vs. after temperature scaling [23]. Entries are mean (std across seeds), to three decimals. Baseline is model-invariant and shown once. ↓ indicates improvement (lower is better), and ↑ indicates degradation. Bold marks the best (minimum) within each tool-LLM block for each metric column.

Tool LLM	Method	NLL		ECE		Brier	
		pre	cal	pre	cal	pre	cal
None	Baseline	1.030 (0.000)	0.706↓ (0.000)	0.114 (0.000)	0.299↑ (0.000)	0.408 (0.000)	0.256↓ (0.000)
Gemma 7B	H1	0.488 (0.055)	0.465↓ (0.039)	**0.057** (0.018)	0.064↑ (0.021)	0.150 (0.013)	0.148↓ (0.011)
	H2	0.470 (0.040)	0.435↓ (0.026)	0.081 (0.023)	**0.035**↓ (0.009)	0.151 (0.015)	0.139↓ (0.011)
	H3	**0.403** (0.017)	**0.411**↑ (0.018)	0.057 (0.020)	0.059↑ (0.006)	**0.128** (0.008)	**0.127**↓ (0.006)
GPT-OSS 20B	H1	0.437 (0.048)	0.427↓ (0.038)	0.045 (0.024)	0.037↓ (0.019)	0.144 (0.020)	0.139↓ (0.015)
	H2	**0.403** (0.022)	**0.400**↓ (0.018)	**0.038** (0.016)	**0.031**↓ (0.009)	**0.128** (0.009)	0.127↓ (0.006)
	H3	0.411 (0.027)	0.415↑ (0.008)	0.060 (0.023)	0.055↓ (0.005)	0.130 (0.010)	**0.125**↓ (0.003)
Qwen3 30B A3B Instruct	H1	**0.429** (0.028)	0.423↓ (0.028)	**0.040** (0.016)	**0.033**↓ (0.011)	**0.139** (0.012)	0.137↓ (0.011)
	H2	0.446 (0.049)	0.429↓ (0.036)	0.067 (0.044)	0.041↓ (0.019)	0.147 (0.020)	0.139↓ (0.014)
	H3	0.447 (0.015)	**0.410**↓ (0.030)	0.092 (0.026)	0.053↓ (0.008)	0.145 (0.006)	**0.129**↓ (0.010)