

LLM Performance Predictors: Learning When to Escalate in Hybrid Human-AI Moderation Systems

Or Bachar

Zefr

Los Angeles, United States
Reichman University
Herzliya, Israel
or.bachar@post.runi.ac.il

Or Levi

Zefr

Los Angeles, United States
or.levi@zefr.com

Sardhendu Mishra

Zefr

Los Angeles, United States
sardhendu.mishra@zefr.com

Adi Levi

Zefr

Los Angeles, United States
adi.levi@zefr.com

Manpreet Singh Minhas

Zefr

Los Angeles, United States
manpreet.minhas@zefr.com

Justin Miller

Zefr

Los Angeles, United States
justin.miller@zefr.com

Omer Ben-Porat

Technion

Haifa, Israel

omerbp@technion.ac.il

Eilon Sheerit

Reichman University

Herzliya, Israel

eilon.sheerit@post.runi.ac.il

Jonathan Morra

Zefr

Los Angeles, United States
jon.morra@zefr.com

ABSTRACT

As LLMs are increasingly integrated into human-in-the-loop content moderation systems, a central challenge is deciding when their outputs can be trusted versus when escalation for human review is preferable. We propose a novel framework for supervised LLM uncertainty quantification, learning a dedicated meta-model based on LLM Performance Predictors (LPPs) derived from LLM outputs: log-probabilities, entropy, and novel uncertainty attribution indicators. We demonstrate that our method enables cost-aware selective classification in real-world human-AI workflows: escalating high-risk cases while automating the rest. Experiments across state-of-the-art LLMs, including both off-the-shelf (Gemini, GPT) and open-source (Llama, Qwen), on multimodal and multilingual moderation tasks, show significant improvements over existing uncertainty estimators in accuracy-cost trade-offs. Beyond uncertainty estimation, the LPPs enhance explainability by providing new insights into failure conditions (e.g., ambiguous content vs. under-specified policy). This work establishes a principled framework for uncertainty-aware, scalable, and responsible human-AI moderation workflows.

KEYWORDS

Large Language Models, Uncertainty Quantification, Human-AI Collaboration, Content Moderation

ACM Reference Format:

Or Bachar, Or Levi, Sardhendu Mishra, Adi Levi, Manpreet Singh Minhas, Justin Miller, Omer Ben-Porat, Eilon Sheerit, and Jonathan Morra. 2026. LLM Performance Predictors: Learning When to Escalate in Hybrid Human-AI Moderation Systems. In *Proc. of the 25th International Conference on*

Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 10 pages. <https://doi.org/10.65109/KWGX1235>

1 INTRODUCTION

Human-AI collaboration is at the heart of modern content moderation, which serves as a cornerstone of online trust and safety. Effective moderation is essential for protecting individuals from harmful or misleading content, and the pursuit of scalable and responsible approaches contributes to a safer online environment. To meet this need, moderation systems must balance the nuanced accuracy of human experts with the scalability and cost-efficiency of AI. The rapid growth of user-generated content has made reliance solely on manual review infeasible, motivating the growing deployment of Large Language Models (LLMs) into moderation workflows. When LLMs are used in human-in-the-loop workflows, their utility depends not only on their accuracy, but also on their ability to signal when they are likely to make an incorrect decision. The effectiveness of such pipelines depends on a single decision that repeats billions of times per day: should the system trust the LLM's judgment or escalate to a human reviewer?

Accordingly, we focus on two research questions:

RQ1: Can we accurately predict when LLMs are likely to be incorrect in moderation tasks?

RQ2: Can selective escalation based on these predictions reduce the overall cost in human-AI moderation workflows?

This trust-or-escalate decision is, at its core, an uncertainty estimation problem. Recent work [6, 9, 25, 33, 40, 47] has explored a variety of signals for estimating LLM uncertainty. Approaches range from token-level entropy and log-probability thresholds [20, 22] to ensemble-based sampling [9] and explicit self-expressed confidence [24, 32, 44, 46].

In the realm of content moderation, current LLM uncertainty metrics fail to capture the economic value of moderation failure. To address this, we formulate the decision as a cost minimization



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). <https://doi.org/10.65109/KWGX1235>

problem, balancing between the cost of misclassification and the cost of human review. Moreover, previous work [22, 40] showed that no single universal uncertainty metric has been proven to be reliable across tasks and models. These findings motivate the pursuit of a robust approach that can perform reliably at scale in high-stakes domains such as content moderation.

To this end, we introduce a novel framework that builds on this foundation, enabling a unified approach to selective escalation. Unlike prior supervised uncertainty estimation approaches that typically rely on single or homogeneous signals [17, 22, 33], or learning-to-defer methods that optimize deferral without uncertainty attribution [11, 37], we fuse diverse gray- and black-box output signals and incorporate moderation-specific abstention with attribution, producing more comprehensive and explainable LLM Performance Predictors (LPPs). The approach is inspired by Query Performance Prediction (QPP) [4, 16, 41–43, 52] in Information Retrieval, which studies how systems can learn to anticipate the success of individual queries. QPPs are commonly divided into pre-retrieval predictors and post-retrieval predictors. In this work, we focus on post-generation predictors, which are scarcer in the literature [20], due to strong empirical predictive power, as they leverage richer evidence from the actual model outputs and broad applicability for both open and closed models.

Uncertainty estimators for LLMs are often categorized by the level of access to model internals: white-box approaches leverage hidden states and gradients, gray-box approaches use output-side signals such as log-probabilities, and black-box approaches rely only on final predictions or agreement across multiple generations. In this work, we focus on gray- and black-box features, as they strike a practical balance between richness of signal, scale, and applicability across both off-the-shelf and open-source LLMs.

Beyond estimating the level of uncertainty, moderation teams also need to know the reason for the uncertainty. We therefore introduce novel moderation-oriented uncertainty attribution indicators that distinguish evidence deficits (e.g. missing transcript context, visually ambiguous frames, cross-lingual inconsistencies) from policy gaps (e.g., edge cases not covered by moderation guidelines, conflicting definitions across regions). This enables practical routing: “Tough Calls” (aleatoric) are routed to senior reviewers, while “Policy Gaps” (epistemic) trigger policy updates or model retraining with active learning. In doing so, LPPs do not merely indicate uncertainty; they also shorten the policy-improvement loop and focus scarce human effort, thereby minimizing the cost of the entire moderation system.

We evaluate the LPPs across multiple families of LLMs, including off-the-shelf (Gemini, GPT) and open-source (Llama, Qwen), on two moderation datasets, featuring multimodal and multilingual content across multiple risk categories: hate, violence, and more. Across settings, LPPs consistently improve the accuracy and cost frontiers relative to each existing uncertainty estimator. We release reproducibility code alongside the paper to enable practitioners to build upon this framework and facilitate future research on LLM Performance Prediction.¹ More broadly, these findings motivate treating trust calibration in moderation as a learned, cost-aware

prediction problem, one that can be implemented without access to model internals.

Contributions. (1) We propose a novel framework for estimating LLM uncertainty, learning a meta-model based on LLM Performance Predictors (LPPs) derived from the LLM’s output; (2) we demonstrate the merits of our method for selective escalation in a real-world hybrid human-AI workflow for content moderation compared to existing uncertainty estimators; and (3) we introduce novel moderation-oriented uncertainty attribution indicators and show their contribution to explainability, offering insights into when and why LLMs fail, thereby improving learning efficiency in human-AI workflows. Due to space limitations, we defer the full mathematical specifications of our LPPs, the cost-ratio sensitivity analysis, prompt templates, and detailed inference configurations to the extended version of the paper [2].

2 RELATED WORK

Content moderation refers to the process of detecting user (e.g., [12, 27, 29]) and AI (e.g., [10, 34, 49]) generated content that violates laws or policies (cf. [12, 13]). The rapid growth of user-generated content, coupled with the vast and evolving range of moderation topics across tasks and use cases, renders manual review impractical, particularly in production environments. This situation drives the increased adoption of unified LLM-based systems for scalable moderation [19, 27, 29, 31]. Recent surveys highlight that uncertainty quantification is a central challenge for deploying such systems reliably at scale [40], and new methods emphasize efficient estimation techniques suitable for production use [47].

The use of LLMs in production moderation systems critically depends on trustworthy confidence estimates, as well-calibrated confidence scores enable safe automation (e.g., selective deferral to human reviewers [23, 37]). There is a large body of work on uncertainty quantification in LLMs [3, 6, 9, 22, 24–26, 28, 32, 33, 35, 40, 44–47, 51]. Prior work has examined verbalized confidence [24, 32, 44, 46], but empirical studies consistently find such methods poorly calibrated and highly sensitive to prompting and model type [28, 46]. In addition, multi-sample methods are computationally expensive at production scale. These findings demonstrate that a single, universal uncertainty metric is unreliable and motivate a more robust, multifaceted approach for model failure prediction. To address this, alternative strategies have been explored, including logit-based calibration [22], reflection or self-consistency signals [35], and reinforcement prompt learning [51]. Some of them, particularly logit-based information, are used in this work.

In this work, we propose a novel framework that leverages a diverse set of predictors to improve both calibration and cost-aware decision-making in moderation. Our feature set fuses logit-based features shown to be effective for textual calibration [22] with additional predictors—including multimodal and uncertainty-attribution signals—that extend applicability beyond text-only settings. Our framework is evaluated across diverse closed- and open-source LLMs on two complementary moderation datasets, including a multimodal benchmark [31], and employs a learned classifier to produce calibrated confidence estimates.

Another line of work closely related to ours is query performance predictors (QPP) for ad hoc document retrieval [4]. These methods

¹<https://github.com/ZEFR-INC/lpp-research>

were devised to predict the effectiveness of search results in the absence of relevance judgments. It is common to divide them into two groups: (i) pre-retrieval predictors, which operate prior to retrieval time and often utilize corpus statistics [16], and (ii) post-retrieval predictors, which leverage information from the retrieved list [41–43, 52]. Recently, and with the advance of LLMs, a new category has emerged: post-generation predictors, which utilize information extracted after LLM generation (e.g., next-token distribution [20]). Pre-retrieval, post-retrieval, and post-generation predictors have also been used to estimate the effectiveness of RAG-based LLM systems [20]. Motivated by this line of work, we incorporate both post-retrieval and post-generation predictors into our feature set, extending them with moderation-oriented attribution features for cost-aware escalation.

3 METHOD

We address the core challenge of uncertainty-aware human-AI collaboration through a four-stage framework. Given the increasingly critical role of LLMs in high-stakes content moderation, we propose a supervised approach to uncertainty quantification that learns when to trust LLM outputs versus escalating for human review. This directly addresses **RQ1** (can we accurately predict LLM errors?) and **RQ2** (can selective escalation reduce overall human-AI moderation costs?).

As illustrated in Figure 1, our framework proceeds in four stages: (1) **Base LLM Inference**: multimodal content is processed by a base LLM using structured prompts that enforce deterministic, token-aligned outputs; (2) **Integer-Token Output Schema**: to ensure consistent probability extraction across diverse base LLMs, we designed prompts that constrain outputs to single integer tokens: 0 = “no”, 1 = “yes”, 2 = “inconclusive_evidence” (aleatoric uncertainty), and 3 = “inconclusive_definition” (epistemic uncertainty), which are validated against a fixed schema; invalid (malformed) outputs are retried deterministically until a valid integer is obtained (≤ 3 retries). These tokens are the basis for uncertainty feature computations [8, 15, 25, 39].² Reasoning tokens are handled separately: they remain free-form and are used only for deriving complementary reasoning-based features (e.g., perplexity, per-token entropy), not for outcome probability computation [14, 47]; (3) **LPP Feature Extraction**: a comprehensive set of LPPs is computed from model outputs, leveraging *gray-box* access (token-level log-probabilities, entropy, and reasoning-path statistics) together with *black-box* compatible features (Verbalized Confidence and Uncertainty Attribution Indicators; see Section 3.1); (4a) **Meta-Model Training**: a Ridge Regression classifier is trained on LPPs to predict LLM correctness, providing calibrated uncertainty estimates [18]; and (4b) **Cost-Aware Routing**: a threshold-based policy decides whether to trust the LLM prediction or escalate to human review, optimizing operational costs.

This architecture is inspired by Query Performance Prediction (QPP) in Information Retrieval [4], adapted to the agent-based setting where a meta-classifier acts as a gating agent, coordinating between an LLM agent and human reviewers.

²Prior work has also proposed an alternative strategy of adding a dedicated uncertainty token (e.g., “[IDK]”), enabling the model to explicitly allocate probability mass to “I don’t know” rather than forcing a distribution over fixed labels [6].

3.1 LLM Performance Predictors (LPPs)

Our LPPs constitute a comprehensive feature set primarily extracted via **gray-box access**, which requires token-level log-probabilities and structured outputs. The feature set also incorporates black-box compatible features (Verbalized Confidence and Uncertainty Attribution Indicators) derived solely from the structured output. This strategy ensures broad applicability, as all evaluated base LLMs provide the necessary gray-box and black-box signals. The combination of gray-box and black-box features strikes a practical balance between information richness and scalability, avoiding reliance on internal activations (white-box) while providing richer signals than output-only methods. Recent analyses have shown that probabilistic confidence (derived from token probabilities) and verbalized confidence capture complementary aspects of model uncertainty: the former is more accurate but requires threshold calibration, whereas the latter provides a reasonable signal without additional setup [38].

LPPs span four families summarized in Table 1; their full mathematical specifications are provided in the extended version [2].

Table 1: Overview of LPP Feature Families. Each category captures a distinct aspect of model uncertainty and is derived from gray-box or black-box signals.

Feature Family	Signal Source & Intuition	Example Features
Post-Hoc Classification Uncertainty (Gray-Box)	Confidence at the final decision boundary, derived from token log-probabilities over discrete classification outcomes.	Max Softmax Probability (MSP), Top-5 Entropy (Entropy), Top-2 Probability Margin (Top-2 Margin)
Internal Generation Uncertainty (Gray-Box; excluded from reported results)	Signals from intermediate Chain-of-Thought (CoT) reasoning tokens.	Perplexity (Natural Base), Sequence Negative Log-Likelihood, Mean Token Entropy, Token Entropy Quantiles
Verbalized Confidence (Black-Box)	Explicit, self-reported confidence stated by the LLM in structured natural language outputs.	Reported Confidence (Scalar), Confidence Bands (One-Hot; VL/L/M/H/VH)
Uncertainty Attribution Indicators (Black-Box)	Uncertainty attribution indicators that distinguish aleatoric uncertainty from epistemic uncertainty.	Evidence-Deficit Indicator, Policy-Gap Indicator

Post-Hoc Classification Uncertainty. Features derived from the probability distribution over outcome tokens, computed over the top- k most probable tokens ($k = 5$), including Top-5 Entropy $H(\hat{p})$, Normalized Top-5 Entropy, Effective Choices ($2^{H(\hat{p})}$), Max Softmax Probability, and Top-2 Probability Margin [1, 9, 25, 26, 40, 45].

Internal Generation Uncertainty. We implemented Chain-of-Thought (CoT) prompting solely to extract reasoning-sequence features (e.g., Perplexity, Sequence Negative Log-Likelihood, Mean Token Entropy, Token Entropy Quantiles, and Token Probability

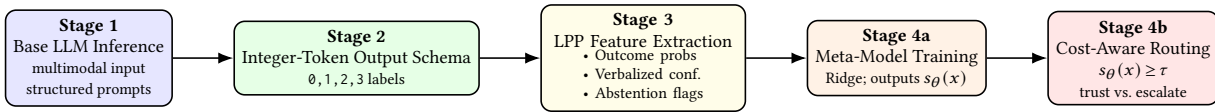


Figure 1: A four-stage framework: (1) Base LLM Inference; (2) Integer-Token Output Schema; (3) LPP Feature Extraction; (4a) Meta-Model Training to estimate $s_\theta(x)$; (4b) Cost-Aware Routing to *trust* or *escalate*.

Quantiles) [1, 14, 47]. In our content-moderation setting, CoT inflated confidence and harmed calibration; therefore, these features are documented for completeness but excluded from the reported results in this work.

Verbalized Confidence. Self-reported confidence extracted from structured outputs, including a scalar confidence score $\hat{c} \in [0, 100]$ (normalized to $[0, 1]$) and confidence bands (Very Low, Low, Medium, High, Very High) encoded as one-hot features [24, 38, 40, 44, 46, 48]. **Uncertainty Attribution Indicators.** Binary indicators capturing cases in which the LLM emits *inconclusive_evidence*, signaling aleatoric uncertainty due to missing context or ambiguous frames, or *inconclusive_definition*, indicating epistemic uncertainty arising from policy gaps or edge cases. These novel moderation-oriented features support interpretable routing [26]: evidence deficient cases are escalated to senior reviewers, while policy-ambiguous cases trigger guideline updates or active learning cycles.

3.2 Prompting

We designed four prompt templates: text-only and multimodal, each with two variants, a direct-answer version (without explicit reasoning steps) and a Chain-of-Thought (CoT) version. All prompts required a structured JSON output with (i) a single integer label (0–3), (ii) an optional reasoning field (populated only in the CoT variant), and (iii) any self-reported confidence. The integer-coded schema aligns labels with token-level log-probabilities. For completeness, we implemented both variants; unless otherwise stated, all results reported in Section 5 use the direct-answer variant.

3.3 Cost-Aware Escalation Policy

The meta-model output is a score $s_\theta(x) \in [0, 1]$ representing the probability that the base LLM is correct. To operationalize this into a binary trust-or-escalate decision, we employ a **cost-aware threshold policy**. A prediction is trusted if $s_\theta(x) \geq \tau$; otherwise, it is escalated to human review [7]. Let c_{mis} denote the cost of an undetected misclassification (allowing an LLM error into production) and c_{rev} denote the cost of human review. The expected cost, relative to a baseline of always trusting the LLM, is:

$$C(\tau) = c_{\text{mis}} \cdot \text{FP} + (c_{\text{rev}} - c_{\text{mis}}) \cdot \text{TN} + c_{\text{rev}} \cdot \text{FN},$$

where FP, TN, FN are counts of selector decisions (trust/escalate) against LLM correctness (correct/incorrect). Here:

- **TP (True Positives):** Trusting a correct LLM prediction \Rightarrow no costs.
- **FP (False Positives):** Trusting an incorrect LLM prediction \Rightarrow misclassification cost (error is missed).
- **TN (True Negatives):** Escalating an incorrect LLM prediction \Rightarrow review cost offset by avoided misclassification cost (error is caught).

- **FN (False Negatives):** Escalating a correct LLM prediction \Rightarrow unnecessary review cost.

The cost of human reviewers was calculated by multiplying the total review time by their hourly rate (USD). The cost of misclassification was estimated based on projected business losses, specifically the reduction in customer lifetime value caused by churn due to undetected errors. Based on these estimates, we set the baseline cost ratio $r := c_{\text{rev}}/c_{\text{mis}} \approx 0.64$ for optimizing the operating threshold τ^* . Varying r over $[0.4, 0.9]$ changes absolute costs but yields consistent qualitative conclusions and cost-accuracy trends. The operating threshold τ^* is optimized on a held-out validation set (20% of the training data) by sweeping $\tau \in [0.35, 0.70]$, a practical range within which the optimum consistently lies near the midpoint (0.5). The value minimizing $C(\tau)$ is then fixed and applied to the held-out test set for evaluation.

3.4 Meta-Model Training

Our meta-model is a supervised binary classifier trained to predict whether the base LLM is correct ($z = 1$) or incorrect ($z = 0$) given the extracted LPPs. If the base LLM outputs an inconclusive label (2/3), we re-run the same prompt with a forced yes/no constraint (0/1 only) to define z , while retaining the original inconclusive type as an attribution feature. Based on extensive preliminary experiments, we selected **Ridge Regression**, whose output scores were converted into probabilities using either sigmoid or isotonic calibration via scikit-learn’s `CalibratedClassifierCV`. This approach was chosen for the following reasons:

Calibration and Robustness. Ridge Regression’s L_2 regularization provides crucial protection against multicollinearity, expected within the LPP feature set where signals like entropy and max probability are inherently correlated. This regularization prevents overfitting and improves out-of-sample calibration, critical for reliable uncertainty quantification in production systems.

Interpretability. Ridge coefficients can be inspected to understand which LPPs are most predictive of errors, supporting explainability requirements in high-stakes moderation workflows.

Handling Class Imbalance. In the Multimodal Moderation Dataset (1,500 samples), base LLM predictions were correct in roughly 80–90% of cases (10–20% errors), resulting in a strong class imbalance. To address this, we downsampled the majority class. We applied a hybrid undersampling strategy combining Tomek Links [30] for boundary cleaning with random undersampling of the remaining majority samples. This reduces redundancy while maintaining representative class ratios, and ensures that rare abstention cases ($< 5\%$ of data) are always retained. We also tested oversampling methods such as SMOTE [5], which yielded comparable performance; for simplicity and reproducibility, we report results using the Tomek+random undersampling approach. We also tuned

Ridge’s `class_weight` via grid search, testing ratios informed by the cost structure: $w_0/w_1 \approx c_{rev}/c_{mis} \approx 0.64$. After downsampling and stratified train–test splitting, our typical training set contains ~ 800 samples and the test set ~ 300 samples for the Multimodal Moderation Dataset, and $\sim 3,500$ training and ~ 900 test samples for the OpenAI Moderation Dataset, with both classes well represented. To ensure fair comparison across LLMs under class imbalance, we fixed the number of negative (error) cases in the test sets: 45 for the Multimodal Moderation Dataset and 135 for the OpenAI Moderation Dataset. This controlled evaluation design mitigates variability due to scarcity of negative samples while preserving comparability across models.

Cross-Validation and Hyperparameter Search. We employ stratified 3-fold cross-validation and a comprehensive grid search on the training portion of the data (after holding out a validation split for threshold selection) over:

- `alpha` $\in \{0.1, 1.0, 10.0, 100.0\}$ (regularization strength)
- `solver` $\in \{\text{auto}, \text{lsqr}\}$
- `tol` $\in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ (convergence tolerance)
- `max_iter` $\in \{1000, 2000, 3000\}$
- `class_weight` $\in \{\text{ratios around } c_{rev}/c_{mis}, \text{“balanced”}\}$
- `method` $\in \{\text{sigmoid}, \text{isotonic}\}$

Models are selected based on F1-score on the minority class (errors), as this metric balances precision and recall for the operationally critical decision of identifying when the LLM is wrong.

Formally, each sample is labeled with a binary correctness indicator $z_i \in \{0, 1\}$, equal to 1 if the LLM prediction matches the ground truth and 0 otherwise. The ridge regression objective is

$$\min_{w,b} \frac{1}{N} \sum_{i=1}^N (z_i - (w^\top f(x_i) + b))^2 + \lambda \|w\|_2^2,$$

where $f(x_i) \in \mathbb{R}^d$ is the feature vector, $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are model parameters, and $\lambda \geq 0$ controls L_2 regularization. This yields a calibrated linear predictor of correctness probability, with isotonic or Platt scaling to map scores into $[0, 1]$. Inference settings are described in Section 4.1.

4 EXPERIMENTS

We evaluate the proposed LPPs framework in the context of cost-aware selective escalation for human-AI content moderation. From a multi-agent systems perspective, our experiments study the coordination dynamics between autonomous LLM agents and human reviewers, where the LPP meta-model serves as a decision-making intermediary that allocates tasks under cost constraints.

4.1 Experimental Setup

We evaluate several LLMs: proprietary (gpt-4o-mini, gpt-4o, gpt-4.1-mini, gemini-2.5-flash-lite, gemini-2.0-flash-lite, gemini-2.0-flash-001) and open-source (QWEN3-14B, QWEN3-32B, LLAMA3.2-11B).³ Following Section 3, we instantiate LPPs using ridge regression trained on heterogeneous feature families derived from intrinsic model signals (e.g., token-level probabilities, self-consistency markers) and post-hoc signals (e.g., calibration-based scores). A summary

³We considered including safety-specialized models such as Llama Guard [21], but defer discussion of these to Section 5.5.

of feature families is provided in Table 1. Hyperparameters for baselines and the meta-model are tuned via nested cross-validation, with stratified sampling. To address class imbalance, we apply downsampling as described in Section 3.4, and evaluate final performance on a held-out test set using the optimized threshold.

Additional Experimental Details. We use deterministic decoding parameters (temperature=0, top-p=1) with a fixed random seed (random_state=42) for data splitting and cross-validation reproducibility. For each generated token, we request log-probabilities of the *top-20* candidate tokens. From these, we compute two feature families:

- (1) **Unfiltered top-5 features:** standard uncertainty metrics computed over the five most probable tokens.
- (2) **Filtered features:** log-probabilities restricted to the four valid class labels (0–3).

This dual representation captures both the model’s general uncertainty landscape and its calibrated confidence over the task’s decision space. CoT perplexity is computed with natural logarithms. Cross-validation uses StratifiedKFold with $K = 3$ splits.

4.2 Datasets

We use two complementary datasets: (i) the **OpenAI Moderation Dataset** [36], which contains 1,680 English text samples annotated across multiple moderation categories (hate, self-harm, sexual, and violence, with fine-grained subcategories such as hate/threatening, sexual/minors, and violence/graphic). Since the dataset is multi-label, we evaluate our model at the level of *text–label pairs*: each annotation is treated as a separate prediction instance. Thus, while the corpus comprises 1,680 source texts, the effective number of evaluation instances is larger, reflecting the total set of text–label assignments. This framing allows a more precise assessment of category-specific moderation performance; (ii) a **Multimodal Moderation Dataset** [31] of 1,500 short-form videos spanning three categories: Death, Injury and Military Conflict; Drugs, Alcohol and Tobacco; and Kids content, across 12 languages and four modalities (text, thumbnail, transcript, video/frame). Together, these datasets (Table 2) enable evaluation of both text-only and multimodal moderation scenarios.

Table 2: Summary of datasets.

Dataset	Size	Languages / Modalities	Categories
OpenAI Moderation [36]	1,680 texts	English / Text	Hate, Self-harm, Sexual, Violence (with subcategories)
Multimodal Moderation [31]	1,500 videos	12 languages / Text, Thumbnail, Transcript, Video/Frames	DIMC (Death/Injury/Military Conflict), DAT (Drugs/Alcohol/Tobacco), Kids

4.3 Baselines

We compare our LPP-based meta-model against widely used uncertainty heuristics:

- **MSP** [11, 17].
- **Top-2 Margin** [15].
- **Entropy** [3, 26].

These represent the strongest post-hoc uncertainty signals. We also include a cost-insensitive *always-trust* baseline.

4.4 Evaluation Metrics

To address **RQ1**, we evaluate error prediction using:

- **AUC-ROC**: area under the receiver operating characteristic curve, measuring ranking quality of correct vs. incorrect predictions across thresholds.
- **F1-Score**: harmonic mean of precision and recall, balancing false positives and false negatives.
- **Macro-F1**: F1 macro-averaged over correctness labels (correct vs. incorrect), giving equal weight to error and non-error cases.

To address **RQ2**, we evaluate:

- **Expected Cost** $C(\tau)$, as defined in Section 3.3.
- **Escalation Rate**, defined as the fraction of items routed to human review:

$$\text{Escalation Rate} = \frac{TN + FN}{TP + FP + TN + FN}.$$

Following best practices in selective prediction [11, 17], we report both the *escalation rate* (percentage of items escalated) and the *absolute number of escalations*.

5 RESULTS

Our meta-model outperforms standard uncertainty estimators in predictive accuracy and cost-aware decision-making across datasets and LLM families. Although we implemented CoT variants, all reported results use direct-answer prompting, as CoT increased confidence without improving calibration.

5.1 Main Performance Benchmarks: Predictive Accuracy (RQ1)

Tables 3 and 4 report error prediction metrics (F1, AUC-ROC, Macro-F1) across nine LLMs on the OpenAI Moderation Dataset [36] and the Multimodal Moderation Dataset [31]. The meta-model consistently exceeds the MSP, Top-2 Margin, and Entropy baselines.

Text-Only (OpenAI Moderation Dataset). The meta-model improves ranking and class balance across models. Relative to the strongest baseline in each row, for **gpt-4o-mini**, F1 increases from 81.27% to 94.14%, AUC-ROC from 83.55% to 93.46%, and Macro-F1 from 64.96% to 82.31%. Similar improvements are observed for **gemini-2.5-flash-lite**, where F1 increases from 82.70% to 89.11%, AUC-ROC from 82.22% to 87.58%, and Macro-F1 from 66.51% to 74.00%, and for **gemini-2.0-flash-lite**, where F1 increases from 88.73% to 93.55%, AUC-ROC from 85.66% to 89.76%, and Macro-F1 from 74.52% to 81.39%. For the stronger-performing models **gpt-4.1-mini** and **gpt-4o**, F1 increases from 88.79% to 91.93% and from 84.41% to 91.35%, respectively. Entries for **LLAMA3.2-11B** and **QWEN3-32B** are omitted from the OpenAI Moderation Dataset results. The former could not be reliably evaluated due to challenges in parsing its outputs across the dataset’s taxonomy, while the latter

required compute resources beyond the scope of our evaluation. These omissions do not affect the aggregate trends reported.

Multimodal (Multimodal Moderation Dataset). On the harder Multimodal Moderation Dataset, effects are more model-dependent. For **gpt-4o-mini**, F1 increases from 85.71% to 87.34%, AUC-ROC from 83.95% to 88.71%, and Macro-F1 from 71.20% to 74.07%. For **gpt-4o**, F1 increases from 88.05% to 91.42%, while Macro-F1 increases from 67.85% to 69.80%. The largest gains are observed for **gemini-2.0-flash-lite**, where F1 increases from 69.85% to 85.47% and Macro-F1 from 52.67% to 61.09%. In contrast, some models exhibit trade-offs: for **gpt-4.1-mini**, Macro-F1 decreases from 72.39% to 69.02%, and for **gemini-2.5-flash-lite**, Macro-F1 decreases from 64.77% to 57.59%. In several cases (e.g., **QWEN3-14B**, **LLAMA3.2-11B**), F1 decreases despite gains in Macro-F1, indicating calibration challenges and majority-class bias under distribution shift.

5.2 Cost-Aware Escalation: Operational Efficiency (RQ2)

Table 5 reports the expected costs and escalation rates under the cost-aware policy described in Section 3.3. These results address RQ2 by quantifying the economic value of selective escalation in human-AI workflows in production.

Significant Cost Reductions on Text-Only Benchmark. On the OpenAI Moderation Dataset, the meta-model achieves substantial cost savings relative to standard baselines for most evaluated models. For **gpt-4o-mini**, the expected cost decreases from \$132 (best baseline: Top-2 Margin) to \$38 (a 71% reduction), while the number of escalations decreases from 331 to 148. Similarly, for **gemini-2.5-flash-lite**, the expected cost decreases from \$122 to \$77 (a 37% reduction), and escalations decrease from 315 to 227. For **gemini-2.0-flash-lite**, the expected cost decreases from \$74 to \$41 (a 45% reduction), with escalations decreasing from 248 to 162. These patterns indicate that baseline methods (MSP, Top-2 Margin, and Entropy) tend to over-escalate, triggering unnecessary human reviews that inflate operational costs without proportional gains in error prevention. In contrast, the meta-model’s improved calibration enables more targeted escalation, reducing false alarms while preserving error detection.⁴

Generalization to Multimodal Workflows. The Multimodal Moderation Dataset exhibits similar trends, with notable variation across models. For **gpt-4o-mini**, the expected cost decreases from \$37 (best baseline: Top-2 Margin) to \$22 (a 41% reduction), with escalations decreasing from 107 to 80. **gpt-4o** shows a comparable pattern, with the expected cost decreasing from \$42 to \$29 (a 31% reduction) and escalations decreasing from 73 to 38. Notably, for **gemini-2.5-flash-lite**, the expected cost decreases from \$51 to \$40 (a 22% reduction), while the number of escalations decreases from 90 to 20. These results suggest that the meta-model can exploit modality-specific uncertainty signals (e.g., visual ambiguity vs. transcript inconsistency) that single-metric baselines fail to capture. However, not all models benefit uniformly: for **QWEN3-14B** and **LLAMA3.2-11B**, the expected cost remains comparable to the Top-2 Margin baseline, indicating limited gains from meta-modeling on this dataset.

⁴Costs are reported in \$, based on estimated review time and projected business loss per error; exact values are task-dependent, and we report relative comparisons.

Table 3: Baseline vs. meta-model predictive performance on the OpenAI Moderation Dataset. Metrics are reported as percentages. Bold values indicate the best result per metric for each LLM (row-wise).

LLM	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
	F1	AUC-ROC	Macro-F1	F1	AUC-ROC	Macro-F1	F1	AUC-ROC	Macro-F1	F1	AUC-ROC	Macro-F1
gpt-4o-mini	81.27%	83.55%	64.96%	81.27%	83.55%	64.96%	80.55%	83.43%	64.20%	94.14%	93.46%	82.31%
gpt-4o	84.35%	90.34%	71.14%	84.41%	90.33%	70.96%	84.35%	90.48%	71.14%	91.35%	90.64%	76.81%
gpt-4.1-mini	88.79%	91.10%	75.28%	88.21%	91.04%	74.56%	88.71%	91.07%	75.16%	91.93%	91.73%	79.94%
gemini-2.5-flash-lite	81.04%	81.69%	65.67%	82.70%	81.41%	66.51%	80.95%	82.22%	65.58%	89.11%	87.58%	74.00%
gemini-2.0-flash-lite	88.73%	85.57%	74.52%	88.48%	85.66%	74.26%	88.73%	85.52%	74.52%	93.55%	89.76%	81.39%
gemini-2.0-flash-001	85.29%	82.76%	70.39%	85.29%	82.76%	70.39%	85.29%	82.76%	70.39%	89.38%	86.34%	75.36%
QWEN3-14B	90.93%	89.94%	64.25%	90.88%	89.90%	63.54%	91.07%	89.99%	66.52%	90.54%	92.07%	77.20%

Table 4: Baseline vs. meta-model predictive performance on the Multimodal Moderation Dataset. Metrics are reported as percentages. Bold values indicate the best result per metric for each LLM (row-wise).

LLM	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
	F1	AUC-ROC	Macro-F1	F1	AUC-ROC	Macro-F1	F1	AUC-ROC	Macro-F1	F1	AUC-ROC	Macro-F1
gpt-4o-mini	85.71%	83.87%	71.20%	85.71%	83.95%	71.20%	85.12%	83.69%	70.47%	87.34%	88.71%	74.07%
gpt-4o	88.05%	84.97%	67.39%	87.58%	84.72%	66.73%	88.00%	84.98%	67.85%	91.42%	84.43%	69.80%
gpt-4.1-mini	88.84%	84.93%	72.39%	88.84%	84.82%	72.39%	88.84%	84.93%	72.39%	90.98%	84.18%	69.02%
gemini-2.5-flash-lite	84.53%	75.18%	63.87%	85.28%	75.20%	64.77%	83.19%	75.21%	62.81%	90.56%	68.26%	57.59%
gemini-2.0-flash-lite	69.52%	67.34%	52.41%	69.85%	67.29%	52.67%	69.52%	67.37%	52.41%	85.47%	69.37%	61.09%
gemini-2.0-flash-001	91.65%	62.38%	45.82%	91.30%	62.52%	52.17%	91.65%	62.32%	45.82%	90.71%	66.37%	56.47%
QWEN3-14B	91.06%	74.04%	60.04%	91.06%	73.92%	60.04%	90.57%	74.25%	60.44%	86.49%	73.77%	65.97%
QWEN3-32B	87.70%	74.70%	55.10%	87.70%	74.52%	55.10%	87.25%	74.79%	54.60%	86.82%	74.71%	57.69%
LLAMA3.2-11B	86.32%	70.79%	56.02%	86.32%	70.49%	56.02%	86.02%	70.74%	59.68%	79.51%	70.67%	62.52%

Table 5: Cost-aware evaluation. Metrics: Always-Trust cost in \$ (A.T.), expected cost in \$ (Cost), number of escalations (Esc), and escalation rate (Rate). Bold values indicate the lowest expected cost among methods for each LLM (row-wise).

LLM	(a) OpenAI Moderation Dataset												(b) Multimodal Moderation Dataset													
	A.T.	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)			A.T.	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
		Cost	Esc	Rate	Cost	Esc	Rate	Cost	Esc	Rate	Cost	Esc	Rate		Cost	Esc	Rate	Cost	Esc	Rate	Cost	Esc	Rate			
gpt-4o-mini	127	138	339	38%	132	331	37%	138	339	38%	38	148	16%	42	39	110	37%	37	107	36%	39	110	37%	22	80	27%
gpt-4o	127	74	275	31%	75	271	30%	74	275	31%	51	151	17%	42	41	70	23%	42	73	24%	40	72	24%	29	38	13%
gpt-4.1-mini	127	69	258	29%	73	267	30%	69	259	29%	45	212	24%	42	32	85	28%	32	85	28%	32	85	28%	30	40	13%
gemini-2.5-flash-lite	127	128	347	38%	122	315	35%	128	348	39%	77	227	25%	42	53	93	31%	51	90	30%	53	92	31%	40	20	7%
gemini-2.0-flash-lite	127	74	248	28%	75	253	28%	74	248	28%	41	162	18%	42	80	163	54%	80	162	54%	80	163	54%	45	64	21%
gemini-2.0-flash-001	127	97	297	33%	97	297	33%	97	297	33%	69	238	26%	42	35	7	2%	35	7	2%	35	10	3%	34	15	5%
QWEN3-14B	127	103	73	8%	105	70	8%	98	84	9%	56	205	23%	42	36	17	6%	36	65	22%	36	21	7%	36	65	22%
QWEN3-32B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	42	47	35	12%	47	46	15%	49	37	12%	47	46	15%
LLAMA3.2-11B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	42	41	25	8%	39	78	26%	38	33	11%	39	78	26%

5.3 Feature Family Ablations: Decomposing LPP Contributions

To quantify the individual contributions of each LPP family, we conduct systematic ablation studies where the meta-model is trained with one feature family removed. Figures 2 and 3 visualize the resulting cost increases relative to the full model across base LLMs. **Complementary Nature of Feature Families.** Across all settings, removing *any* feature family degrades performance, confirming that the three LPP families (Post-Hoc Classification Uncertainty, Verbalized Confidence, and Uncertainty Attribution Indicators) capture complementary aspects of model behavior. On the OpenAI Moderation Dataset (Figure 2), removing Post-Hoc Classification

Uncertainty features (Entropy, MSP, Top-2 Margin) leads to the largest cost increases for most models, consistent with prior work showing that token-level probabilities are the strongest single uncertainty signal. Removing verbalized confidence also increases costs by roughly 5–15%, as indicated by the cost differences in Figures 2 and 3, suggesting that self-reported confidence adds complementary signals not fully captured by probabilistic features.

The Uncertainty Attribution Indicators, despite being binary, contribute measurably to cost reduction, with especially visible effects for **gemini-2.5-flash-lite** and **gemini-2.0-flash-lite**, where removing them produces clear increases in expected costs. This highlights the value of explicit abstention signals in helping the

meta-model avoid false confidence—cases where the base LLM assigns high probability to an incorrect answer due to policy ambiguity or evidence deficits.

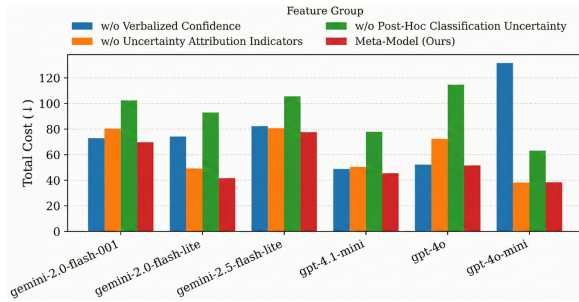


Figure 2: Ablation study on the OpenAI Moderation Dataset. Bars show total cost (\$) when removing one feature family, illustrating each predictor family’s contribution.

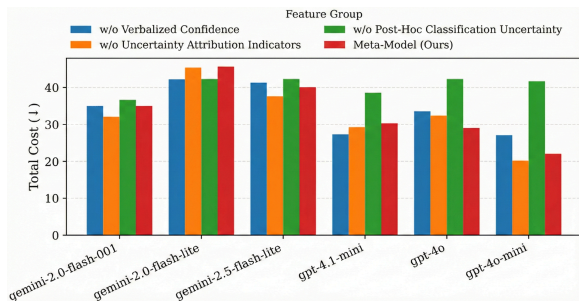


Figure 3: Ablation study on the Multimodal Moderation Dataset. Bars show total cost (\$) when removing one feature family, illustrating each predictor family’s contribution.

5.4 Cross-Dataset Generalization and Robustness

A key question for meta-learning is whether patterns transfer across distributions. Our two-dataset evaluation shows *model-dependent* robustness: on the OpenAI Moderation Dataset (text-only), the meta-model consistently outperforms the baselines, while on the Multimodal Moderation Dataset it improves the performance for some models (e.g., gpt-4o-mini, gpt-4o, gemini-2.0-flash-lite) but is matched or outperformed by others (e.g., Qwen, Llama), reflecting heterogeneity under domain shift.

Absolute performance naturally degrades for some models when moving from the text-only benchmark to the harder multimodal distribution. For example, for **gpt-4.1-mini**, the Macro-F1 score decreases from 79.94% (text-only) to 69.02% (multimodal), illustrating the additional complexity introduced by visual ambiguity, multilingual code-switching, and culturally contingent moderation judgments. Despite these shifts, the meta-model retains gains for multiple models on the Multimodal Moderation Dataset, while revealing cases where baseline methods remain competitive, underscoring that generalization depends on both the base LLM family and the evaluation distribution.

These patterns reinforce the *cost-aware* view: even when accuracy gains are mixed under distribution shift, LPP-based escalation yields savings by directing review to the most error-prone cases.

5.5 Limitations and Open Challenges

Our study demonstrates substantial progress, but several limitations remain. First, the meta-model requires labeled data samples with ground-truth moderation decisions, which introduces annotation costs. Semi-supervised and active learning strategies could reduce this burden. Second, we focus only on *post-response* LPPs extracted after the LLM generates an answer. Incorporating *pre-response* predictors could enable anticipatory escalation and adaptive coordination. Third, although we evaluate across a range of LLM families, all are transformer-based autoregressive models. Whether LPPs generalize to other architectures (e.g., retrieval-augmented or neuro-symbolic systems) remains an open question. Fourth, our cost model assumes fixed review and error penalties, whereas in practice costs vary by severity, reviewer expertise, and platform-specific risk tolerance. Extending escalation to a dynamic resource-allocation problem is a promising direction. Finally, while we extracted Chain-of-Thought (CoT) derived features, we do not advocate CoT prompting in this setting: in our experiments, it inflated confidence and harmed calibration [44, 50].

More broadly, while our experiments focus on content moderation, the LPP framework generalizes to other high-stakes decision settings such as fraud detection, compliance, or medical triage, where abstention signals and cost-sensitive escalation are equally critical. By abstracting uncertainty-aware routing beyond moderation, LPPs can serve as a general-purpose coordination mechanism wherever human expertise must be allocated efficiently under uncertainty. We excluded safety-specialized models such as Llama Guard [21], whose taxonomies do not align with our categories (DIMC: Death, Injury, Military Conflict; DAT: Drugs, Alcohol, Tobacco; Kids). Such models may offer stronger abstention signals or calibration, and hybrid setups with domain-specialized safety models are a natural future direction.

6 CONCLUSIONS

Synthesis: Toward Uncertainty-Aware Multi-Agent Moderation. Our empirical findings collectively establish that supervised LPPs enable a principled, cost-effective approach to human-AI collaboration in content moderation. The meta-model acts as an intelligent gating agent, coordinating between autonomous LLM agents and human reviewers by dynamically allocating tasks based on predicted error likelihood and operational constraints.

The LPP framework also opens research directions at the intersection of uncertainty and multi-agent coordination: (i) hierarchical escalation to reviewers with varying expertise, (ii) federated learning for privacy-preserving training across platforms, and (iii) integration with RLHF to couple uncertainty estimation with model improvement through iterative feedback loops.

Ultimately, effective human-AI collaboration depends not only on accuracy but on systems that *know what they don’t know*. By quantifying and attributing uncertainty, LPPs turn opaque LLM outputs into actionable signals, enabling reviewers to focus their expertise where it matters most.

REFERENCES

- [1] Andrew Gabriel Araujo Correa and Ana Carolina Hermogenes de Matos. 2025. Entropy-Guided Loop: Achieving Reasoning through Uncertainty-Aware Generation. <https://zenodo.org/doi/10.5281/zenodo.16955206>
- [2] Or Bachar, Or Levi, Sardhendu Mishra, Adi Levi, Manpreet Singh Minhas, Justin Miller, Omer Ben-Porat, Eilon Sheerit, and Jonathan Morra. 2026. LLM Performance Predictors: Learning When to Escalate in Hybrid Human-AI Moderation Systems. *arXiv:2601.07006 [cs.AI]* <https://arxiv.org/abs/2601.07006>
- [3] Guy Bar-Shalom, Fabrizio Frasca, Derek Lim, Yoav Gelberg, Yftah Ziser, Ran El-Yaniv, Gal Chechik, and Haggai Maron. 2025. Beyond Next Token Probabilities: Learnable, Fast Detection of Hallucinations and Data Contamination on LLM Output Distributions. <https://arxiv.org/abs/2503.14043>
- [4] David Carmel and Elad Yom-Tov. 2010. Estimating the query difficulty for information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, USA, 911–911.
- [5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. <https://www.jair.org/index.php/jair/article/view/10302>
- [6] Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I Don't Know: Explicit Modeling of Uncertainty with an [IDK] Token. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 10935–10958. <https://doi.org/10.52202/079017-0349>
- [7] Charles Elkan. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (Seattle, WA, USA) (IJCAI'01). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978.
- [8] David Farr, Nico Manzoni, Iain Cruickshank, and Jevin West. 2025. RED-CT: A Systems Design Methodology for Using LLM-labeled Data to Train and Deploy Edge Linguistic Classifiers. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*. Association for Computational Linguistics, Abu Dhabi, UAE, 58–67. <https://aclanthology.org/2025.coling-industry.5/>
- [9] David T. Farr, Lynnette Hui Xian Ng, Iain J. Cruickshank, Nico Manzoni, Nicholas Clark, Kate Starbird, Nathaniel D. Bastian, and Jevin West. 2025. Ensemble-Based Uncertainty Quantification for Reliable Large Language Model Classification in Social Data Applications. *IEEE Access* 13 (2025), 116419–116429.
- [10] Sarah A Fisher, Jeffrey W Howard, and Beatriz Kira. 2024. Moderating synthetic content: The challenge of generative AI. *Philosophy & Technology* 37, 4 (2024), 133.
- [11] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. <https://arxiv.org/abs/1705.08500>
- [12] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
- [13] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
- [14] Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 1059–1075. <https://aclanthology.org/2023.eacl-main.75/>
- [15] Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language Model Cascades: Token-Level Uncertainty And Beyond. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=KgaBScZ4V1>
- [16] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 1419–1420.
- [17] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Hkg4TI9xl>
- [18] Arthur E. Hoerl and Robert W. Kennard. 2000. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 42, 1 (2000), 80–86. <http://www.jstor.org/stable/1271436>
- [19] Tao Huang. 2025. Content moderation by LLM: From accuracy to legitimacy. *Artificial Intelligence Review* 58, 10 (2025), 1–32.
- [20] Oz Huly, David Carmel, and Oren Kurland. 2025. Predicting RAG Performance for Text Completion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1283–1293.
- [21] Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madihan Khabsa. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. *arXiv preprint arXiv:2312.06674* abs/2312.06674 (2023). <https://api.semanticscholar.org/CorpusID:266174345>
- [22] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics* 9 (09 2021), 962–977. https://doi.org/10.1162/tacl_a_00407
- [23] Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: LLM judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370* (2024).
- [24] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).
- [25] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large Language Models Must Be Taught to Know What They Don't Know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=QzvWyggrYB>
- [26] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision?. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17)*. Curran Associates Inc., 5580–5590.
- [27] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-mod: Can large language models assist content moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [28] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664* (2023).
- [29] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AI Conference on Web and Social Media*, Vol. 18. 865–878.
- [30] Wonjae Lee and Kangwon Seo. 2022. Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning. *Pattern Recognition Letters* 170 (2022), 207–214. <https://www.sciencedirect.com/science/article/pii/S2214579622000089>
- [31] Adi Levi, Or Levi, Sardhendu Mishra, and Jonathan Morra. 2025. AI vs. Human Moderators: A Comparative Evaluation of Multimodal LLMs in Content Moderation for Brand Safety. *arXiv preprint arXiv:2508.05527* (2025).
- [32] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3214–3252.
- [33] Linyu Liu, Yu Pan, Xiaocheng Li, and Quanting Chen. 2024. Uncertainty estimation and quantification for LLMs: A simple supervised approach. *arXiv preprint arXiv:2404.15993* (2024).
- [34] Travis Lloyd, Joseph Reagle, and Mor Naaman. 2023. "There Has To Be a Lot That We're Missing": Moderating AI-Generated Content on Reddit. *arXiv preprint arXiv:2311.12702* (2023).
- [35] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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*. 9004–9017.
- [36] Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A Holistic Approach to Undesired Content Detection in the Real World. <https://arxiv.org/abs/2208.03274>
- [37] Hussein Mozannar and David Sontag. 2021. Consistent Estimators for Learning to Defer to an Expert. <https://arxiv.org/abs/2006.01862>
- [38] Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2025. Are Large Language Models More Honest in Their Probabilistic or Verbalized Confidence?. In *Information Retrieval*. Springer Nature Singapore, 124–135.
- [39] Hadas Orgad, Michael Tokar, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. <https://arxiv.org/abs/2410.02707>
- [40] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. <https://arxiv.org/abs/2412.05563>
- [41] Anna Shtok, Oren Kurland, and David Carmel. 2016. Query performance prediction using reference lists. *ACM Transactions on Information Systems (TOIS)* 34, 4 (2016), 1–34.
- [42] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems (TOIS)* 30, 2 (2012), 1–35.
- [43] Yongquan Tao and Shengli Wu. 2014. Query performance prediction by considering score magnitude and variance together. In *Proceedings of the 23rd ACM international conference on information and knowledge management*. 1891–1894.
- [44] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models

- Fine-Tuned with Human Feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- [45] David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. 2023. Explaining Predictive Uncertainty with Information Theoretic Shapley Values. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=6rabAZhCRS>
- [46] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint arXiv:2306.13063* (2023).
- [47] Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. 2024. Efficient and Effective Uncertainty Quantification in LLMs. In *NeurIPS Workshop*. <https://openreview.net/forum?id=QKRLH57ATT>
- [48] Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On Verbalized Confidence Scores for LLMs. <https://arxiv.org/abs/2412.14737>
- [49] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems* 32 (2019).
- [50] Hanlin Zhang, Yi-Fan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, and Sham Kakade. 2024. A Study on the Calibration of In-context Learning. <https://arxiv.org/abs/2312.04021>
- [51] Tianjun Zhang, Xuezi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022. Tempera: Test-time prompting via reinforcement learning. *arXiv preprint arXiv:2211.11890* (2022).
- [52] Yun Zhou and W Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 543–550.