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

Or Bachar*
Zefr
Los Angeles, United States
or.bachar@zefr.com
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 Sheetrit
Reichman University
Herzliya, Israel
eilon.sheetrit@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, Content Moderation, Human–AI Collaboration, Cost-aware Routing

ACM Reference Format:
Or Bachar, Or Levi, Sardhendu Mishra, Adi Levi, Manpreet Singh Minhas, Justin Miller, Omer Ben-Porat, Eilon Sheetrit, 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, 12 pages.

*Also with Reichman University.

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). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

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 [5, 8, 24, 32, 39, 46] has explored a variety of signals for estimating LLM uncertainty. Approaches range from token-level entropy and log-probability thresholds [19, 21] to ensemble-based sampling [8] and explicit self-expressed confidence [23, 31, 43, 45].

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 problem, balancing between the cost of misclassification and the cost of human review. Moreover, previous work [21, 39] 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 [16, 21, 32], or learning-to-defer methods that optimize deferral without uncertainty attribution [10, 36], 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) [3, 15, 40–42, 51] 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 more scarce in the literature [19], 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, 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]

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.

[1] https://github.com/ZEFR-INC/lpp-research

2 Related Work

Content moderation refers to the process of detecting user (e.g., [11, 26, 28]) and AI (e.g., [9, 33, 48]) generated content that violates laws or policies (cf. [11, 12]). 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 [18, 26, 28, 30]. Recent surveys highlight that uncertainty quantification is a central challenge for deploying such systems reliably at scale [39], and new methods emphasize efficient estimation techniques suitable for production use [46].

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 [22, 36]). There is a large body of work on uncertainty quantification in LLMs [2, 5, 8, 21, 23–25, 27, 31, 32, 34, 39, 43–46, 50]. Prior work has examined verbalized confidence [23, 31, 43, 45], but empirical studies consistently find such methods poorly calibrated and highly sensitive to prompting and model type [27, 45]. 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 [21], reflection or self-consistency signals [34], and reinforcement prompt learning [50]. 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 [21] 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 [30], 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 [3]. These methods 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 [15], and (ii) post-retrieval predictors, which leverage information from the retrieved list [40–42, 51]. 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 [19]). Pre-retrieval, post-retrieval, and post-generation predictors have also been used to estimate the effectiveness of RAG-based LLM systems [19]. 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

Our method addresses the core challenge of uncertainty-aware human-AI collaboration through a principled 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 model outputs versus when to escalate for human review. This design directly addresses our research questions: **RQ1** (can we accurately predict LLM errors?) and **RQ2** (can selective escalation based on these predictions reduce the overall cost in human-AI moderation workflows?). 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 [7, 14, 24, 38].[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 [13, 46]; (3) **LPP Feature Extraction**: a comprehensive set of LLM Performance Predictors (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 [17]; 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 [3], adapted to the agent-based setting where a meta-classifier acts as a gating agent, coordinating between an LLM agent and human reviewers.

3.1 LLM Performance Predictors (LPPs)

Our LLM Performance Predictors (LPPs) are a comprehensive feature set primarily extracted via **gray-box access**—requiring token-level log-probabilities and structured outputs. This feature set also incorporates black-box compatible features (Verbalized Confidence and Uncertainty Attribution Indicators) derived solely from the structured text output. This strategy ensures broad applicability: 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 analysis has shown that probabilistic confidence (derived from token probabilities) and verbalized confidence capture complementary aspects of model uncertainty, with the former being more accurate but requiring threshold calibration, while the latter provides reasonable signal without additional setup [37]. LPPs span four families, summarized in Table 1, and the full mathematical specification of all LPPs appears in Table 6 (Appendix A).

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 $H(\tilde{p})$, Normalized Top-5 Entropy, Effective Choices ($2^{H(\tilde{p})}$), Max Softmax Probability, and Top-2 Probability Margin [1, 8, 24, 25, 39, 44].

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 Quantiles) [1, 13, 46]. In our content-moderation setting, CoT inflated confidence and harmed calibration; therefore, these features are documented for completeness but excluded from reported results.

Verbalized Confidence. Self-reported confidence extracted from structured outputs: a scalar reported confidence $\hat{c} \in [0, 100]$ (normalized to $[0, 1]$) and coarse confidence bands (Very Low, Low, Medium, High, Very High) encoded as one-hot features [23, 37, 39, 43, 45, 47].

Uncertainty Attribution Indicators. Binary indicators when the LLM emits inconclusive_evidence (aleatoric uncertainty due to missing context, ambiguous frames, etc.) or inconclusive_definition (epistemic uncertainty due to policy gaps or edge cases). These novel moderation-oriented features enable interpretable routing [25]: evidence deficient cases escalate 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 (no 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 [6]. Let c_{mis} denote the cost of an undetected misclassification (allowing an LLM error into production) and c_{rev} the cost of human review. The expected cost, relative to a baseline of always trusting the LLM, is:

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

[2] Prior 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 [5].

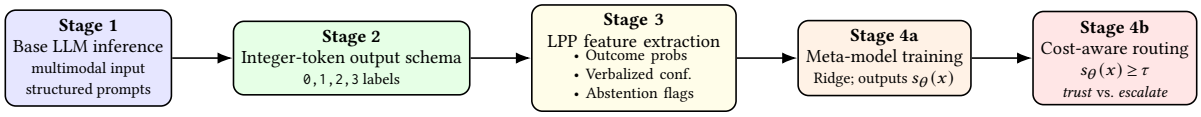

Figure 1: A four-stage framework: (1) Base LLM Inference; (2) Integer-Token Output Schema; (3) LPP Feature Extraction; (4a) Meta-Model Training to estimate $s_\theta(x)$; (4b) Cost-Aware Routing to *trust* or *escalate*.

Table 1: Overview of LPP Feature Families. Each category captures a distinct aspect of model uncertainty and is accessible via gray-box or black-box access. The full mathematical specification of all LPPs appears in Table 6 (Appendix A).

LPP Category	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), Confidence Score
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 signals distinguishing aleatoric uncertainty (evidence deficits) from epistemic uncertainty (policy gaps). Enables interpretable routing.	Inconclusive Flag (Binary), Evidence-Deficit Indicator, Policy-Gap Indicator

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

The cost of human reviewers was calculated by multiplying the total review time by their hourly rate (in $). 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. In our experiments, we set the cost ratio such that $c_{\text{rev}}/c_{\text{mis}} \approx 0.64$, which serves as a reasonable baseline and does not alter relative rankings or cost–accuracy trends under variation. The operating threshold $\tau^\star$ 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. 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.

Gray-Box Applicability. Ridge Regression does not require model internals (hidden states, gradients), ensuring compatibility with both off-the-shelf APIs and open-source LLMs.

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 (1) downsampled the majority class. We applied a hybrid undersampling strategy combining Tomek Links [29] 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 [4], 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: $w_0/w_1 \approx c_{\text{rev}}/c_{\text{mis}} \approx 0.64$. 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 ~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 150 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` $\in$ {auto, lsqr}
- `tol` $\in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ (convergence tolerance)
- `max_iter` $\in \{1000, 2000, 3000\}$
- `class_weight`: 7 configurations including cost-informed ratios and "balanced"
- Calibration `method` $\in$ {sigmoid, 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

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

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 (including deterministic decoding) are described in Section 4.1 and Appendix A.

4 Experiments

We evaluate the proposed LLM Performance Predictor (LPP) 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, LLAMA32-11B). [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 of feature groups 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 Sec. 3.4, and evaluate final performance on a held-out test set using the optimized threshold. The meta-model is trained with Ridge Regression, chosen for its robustness in high-dimensional feature spaces.

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.

[3] We considered including safety-specialized models such as Llama Guard [20], but defer discussion of these to Section 5.5.

Table 2: Summary of datasets.

Dataset	Size	Languages / Modalities	Categories
OpenAI Moderation [35]	1,680 texts	English / Text	Hate, Self-harm, Sexual, Violence (with subcategories)
Multimodal Moderation [30]	1,500 videos	12 languages / Text, Thumbnail, Transcript, Video/Frames	DIMC (Death/Injury/Military Conflict), DAT (Drugs/Alcohol/Tobacco), Kids

4.2 Datasets

We use two complementary datasets: (i) the **OpenAI Moderation dataset** [35], 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** [30] 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 enable evaluation of both text-only and multimodal moderation scenarios.

4.3 Baselines

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

- **MSP** [10, 16].
- **Top-2 Margin** [14].
- **Entropy** [2, 25].

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 Sec. 3.3
- **Escalation Ratio**, defined as the fraction of items routed to human review:

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

Following best practice in selective prediction [10, 16], we report both the *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 F1, AUC-ROC, and Macro-F1 for error prediction across nine LLMs overall, evaluated on two datasets: OpenAI Moderation Dataset [35] and a Multimodal Moderation Dataset [30]. The meta-model consistently exceeds 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 **LLAMA32-11B** 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**, **LLAMA32-11B**), 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 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 production human–AI workflows.

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 models. For **gpt-4o-mini**, expected cost decreases from \$132 (best baseline: Top-2 Margin) to \$38 (71% reduction), while the number of escalations decreases from 331 to 148. Similarly, for **gemini-2.5-flash-lite**, expected cost decreases from \$122 to \$77 (36% reduction) and escalations from 315 to 227. For **gemini-2.0-flash-lite**, expected cost decreases from \$74 to \$41 (41% reduction), with escalations decreasing from 248 to 162. These patterns indicate that baseline methods (e.g., MSP, 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**, expected cost decreases from \$37 (best baseline: Top-2 Margin) to \$22 (41% reduction), with escalations decreasing from 107 to 80. **gpt-4o** shows a comparable pattern, with expected cost decreasing from \$42 to \$29 (32% reduction) and escalations from 73 to 38. Notably, for **gemini-2.5-flash-lite**, expected cost decreases from \$51 to \$40 (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 **LLAMA32-11B**, expected cost remains comparable to the Top-2 Margin baseline, indicating limited gains from meta-modeling on this dataset.

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 group 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 groups (Post-Hoc Classification Uncertainty, Verbalized Confidence, and Uncertainty Attribution Indicators) capture complementary aspects of model behavior. In Figures 2 and 3, these correspond respectively to *Outcome-Level*, *Verbalized*, and *Inconclusive* feature groups. On the OpenAI Moderation Dataset (Figure 2), removing Post-Hoc Classification Uncertainty features

[4] Costs 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).

	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
LLM	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).

	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
LLM	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%**
LLAMA32-11B	**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 (A.T.) cost ($), expected cost ($), escalations (count), and escalation ratio (%). Bold values indicate the lowest expected cost among methods for each LLM (row-wise).

(a) OpenAI Moderation Dataset

LLM	A.T.	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
		Cost	Esc	Ratio	Cost	Esc	Ratio	Cost	Esc	Ratio	Cost	Esc	Ratio
gpt-4o-mini	127	138	339	38%	132	331	37%	138	339	38%	**38**	148	16%
gpt-4o	127	74	275	31%	75	271	30%	74	275	31%	**51**	151	17%
gpt-4.1-mini	127	69	258	29%	73	267	30%	69	259	29%	**45**	212	24%
gemini-2.5-flash-lite	127	128	347	38%	122	315	35%	128	348	39%	**77**	227	25%
gemini-2.0-flash-lite	127	74	248	28%	75	253	28%	74	248	28%	**41**	162	18%
gemini-2.0-flash-001	127	97	297	33%	97	297	33%	97	297	33%	**69**	238	26%
QWEN3-14B	127	103	73	8%	105	70	8%	98	84	9%	**56**	205	23%
QWEN3-32B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LLAMA32-11B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

(b) Multimodal Moderation Dataset

A.T.	MSP			Top-2 Margin			Entropy			Meta-Model (Ours)		
	Cost	Esc	Ratio	Cost	Esc	Ratio	Cost	Esc	Ratio	Cost	Esc	Ratio
42	39	110	37%	37	107	36%	39	110	37%	**22**	80	27%
42	41	70	23%	42	73	24%	40	72	24%	**29**	38	13%
42	32	85	28%	32	85	28%	32	85	28%	**30**	40	13%
42	53	93	31%	51	90	30%	53	92	31%	**40**	20	7%
42	80	163	54%	80	162	54%	80	163	54%	45	64	21%
42	35	7	2%	35	7	2%	35	10	3%	**34**	15	5%
42	**36**	17	6%	**36**	65	22%	**36**	21	7%	**36**	65	22%
42	47	35	12%	47	46	15%	49	37	12%	47	46	15%
42	41	25	8%	39	78	26%	**38**	33	11%	39	78	26%

(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.

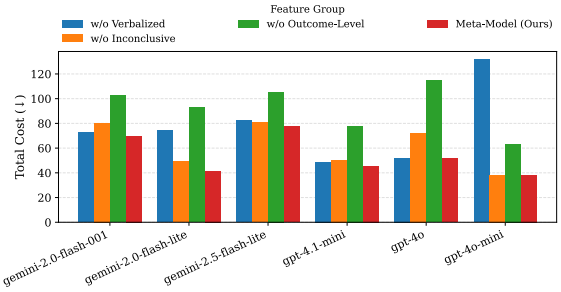

Figure 2: Ablation study on the OpenAI Moderation Dataset. Bars show total cost ($) when removing one feature family, illustrating each predictor group's contribution.

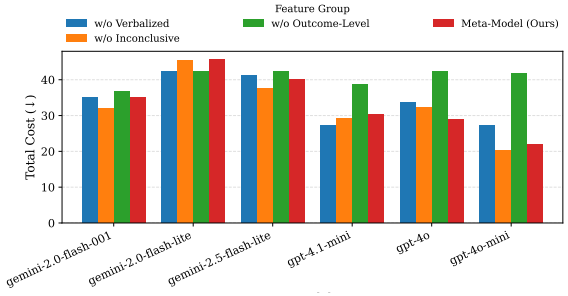

Figure 3: Ablation study on the Multimodal Moderation Dataset. Bars show total cost ($) when removing one feature family, illustrating each predictor group's contribution.

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 post-hoc baselines, while on the Multimodal Moderation Dataset it improves 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%**, 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

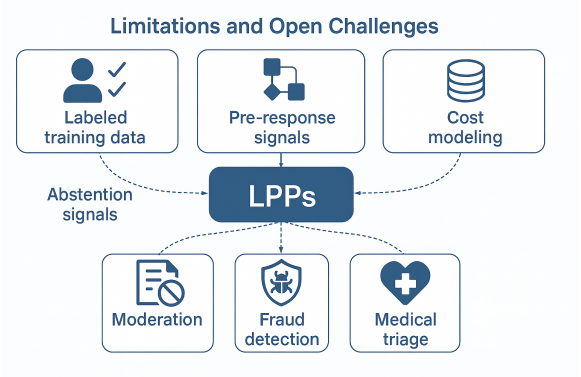

Figure 4: LPPs as a cross-domain uncertainty router. The meta-model routes between automated decisions and human review using uncertainty signals, applicable to moderation, fraud detection, compliance, and medical triage.

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 [43, 49].

More broadly, our experiments focus on content moderation, but the LPP framework is applicable to other high-stakes settings such as fraud detection, compliance, or medical triage, where abstention signals and cost-sensitive escalation are equally critical. Figure 4 illustrates this broader vision: by abstracting uncertainty-aware routing beyond moderation, LPPs can act as a general-purpose coordination mechanism wherever human expertise must be allocated efficiently under uncertainty. We excluded safety-specialized models like Llama Guard [20], 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 Conclusion

Synthesis: Toward Uncertainty-Aware Multi-Agent Moderation. Our empirical findings collectively establish that supervised LLM Performance Predictors 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.

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] 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

[3] 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*. 911–911.

[4] 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

[5] Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I Don't Know: Explicit Modeling of Uncertainty with an [IDK] Token. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=Wc0vlQuoLb

[6] Charles Elkan. 2001. The foundations of cost-sensitive learning *(IJCAI'01)*. Morgan Kaufmann Publishers Inc., 973–978.

[7] David Farr, Nico Manzonelli, 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/

[8] David T. Farr, Lynnette Hui Xian Ng, Iain J. Cruickshank, Nico Manzonelli, 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.

[9] Sarah A Fisher, Jeffrey W Howard, and Beatriz Kira. 2024. Moderating synthetic content: The challenge of generative AI. *Philosophy & Technology* 37, 4 (2024), 133.

[10] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. https://arxiv.org/abs/1705.08500

[11] 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.

[12] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[13] 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/

[14] 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=KgaBScZ4VI

[15] 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.

[16] 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

[17] 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

[18] Tao Huang. 2025. Content moderation by LLM: From accuracy to legitimacy. *Artificial Intelligence Review* 58, 10 (2025), 1–32.

[19] 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.

[20] Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian 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

[21] 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 (2021), 962–977.

[22] 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).

[23] 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).

[24] 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

[25] 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.

[26] 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.

[27] 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).

[28] 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.

[29] 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

[30] 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).

[31] 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.

[32] Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for LLMs: A simple supervised approach. *arXiv preprint arXiv:2404.15993* (2024).

[33] 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).

[34] 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.

[35] 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

[36] Hussein Mozannar and David Sontag. 2021. Consistent Estimators for Learning to Defer to an Expert. https://arxiv.org/abs/2006.01862

[37] 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.

[38] Hadas Orgad, Michael Toker, 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

[39] 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

[40] 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.

[41] 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.

[42] 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 conference on information and knowledge management*. 1891–1894.

[43] 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*.

[44] 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

[45] 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).
[46] 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
[47] Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On Verbalized Confidence Scores for LLMs. https://arxiv.org/abs/2412.14737
[48] 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).
[49] 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
[50] Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022. Tempera: Test-time prompting via reinforcement learning. *arXiv preprint arXiv:2211.11890* (2022).
[51] 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.

A Prompt Templates and Inference Configuration

This appendix documents the schemas, prompt templates, and inference settings used to generate LLM outputs. Placeholders `{{TEXT}}`, `{{TRANSCRIPT}}`, `{{THUMBNAIL}}`, `{{VIDEO\FRAMES}}`, and `{{CONCEPT_DEFINITION}}` were filled dynamically for each query.

A.1 Structured Output Schema

Models were instructed to emit a `JSON` object conforming to a predefined schema with a single integer outcome in: `0` = "no", `1` = "yes", `2` = "inconclusive_evidence", `3` = "inconclusive_definition", which are validated against a fixed schema; invalid (malformed) outputs trigger up to three deterministic retries.

The `ConceptClassification` object contains the final `outcome`, optional `reasoning_steps`, self-reported confidence `p_correct`, and a confidence `band`.

For reproducibility, the overall `LLMClassification` object and `ReasoningStep` are shown below.

```
class LLMClassification(BaseModel):
   classifications: Dict[ConceptEnum, ConceptClassification]

class ConceptClassification(BaseModel):
    outcome: Optional[OutcomeEnum]
    reasoning_steps: Optional[List[ReasoningStep]]
    p_correct: Optional[int]
    band: Optional[BandEnum]

class ReasoningStep(BaseModel):
    step_number: int
    description: str
```

Schema Clarifications. All CoT prompts enforced exactly three `reasoning_steps`. We used regular expressions to normalize the outcome tokens `(yes|no|inconclusive_(definition|evidence))` for consistent parsing. The `p_correct` field was snapped to the nearest multiple of 5 in $[0, 100]$, and the `band` was validated against `{VL, L, M, H, VH}`.

A.2 Representative Prompt Templates

Text-only Baseline (System Prompt).

```
# ROLE AND GOAL
You are a meticulous Chief Marketing Officer (CMO) ...
```

Text-only User Prompt.

```
Please classify the following content:
--- CONTENT START ---
{{TEXT}}
--- CONTENT END ---
```

Multimodal User Prompt.

```
Analyze the following multimodal content.

--- START OF MULTIMODAL CONTENT ---
<VIDEO\FRAMES> {{VIDEO\FRAMES}} </VIDEO\FRAMES>
<THUMBNAIL> {{THUMBNAIL}} </THUMBNAIL>
<TRANSCRIPT> {{TRANSCRIPT}} </TRANSCRIPT>
<CONTENT_TEXT> {{TEXT}} </CONTENT_TEXT>
--- END OF MULTIMODAL CONTENT ---
```

Depending on the base LLM, the visual input was provided either as a video URI or as a set of key frames; the placeholder `{{VIDEO\FRAMES}}` denotes this modality. The CoT variants added a `REASONING FRAMEWORK` section to the system prompt to enforce a step-by-step thought process. The multimodal variants included placeholders for all four modalities and a protocol requiring a review of each.

A.3 Inference Configuration

The following fixed decoding parameters were used for all API calls:

- Decoding: `temperature=0`, `top-p=1`, `n=1`.
- Max output tokens: `8096`.
- Log-probabilities: top-20 per token.

Tokens were segmented into classification and reasoning spans based on the structured output. Assets (images, videos) were passed as Google Cloud Storage URIs, and malformed JSON responses (fewer than 2% of total requests) were retried up to three times.

A.4 Cost-Ratio Sensitivity

We recompute the relative expected cost at the fixed operating point $\tau^\star$ under cost-ratio variation $r = c_{\text{rev}}/c_{\text{mis}} \in [0.4, 0.9]$ using the confusion-derived counts (FP, TN, FN) and the cost definition in §3.3: $C/c_{\text{mis}} = FP + (r-1)\cdot TN + r \cdot FN$. Table 7 reports values at representative ratios $\{0.4, 0.64, 0.9\}$ spanning the swept range.

Table 6: Catalog of LLM Performance Predictors (LPPs). Features are grouped by access (gray-box vs. black-box). A–B: Outcome-level confidence metrics (top-k, schema-filtered). C: Log-probability and confidence margins. D–E: Reasoning-path features from CoT (perplexity, entropy). F: Self-reported confidence. G: Moderation-oriented abstention signals, distinguishing *aleatoric* (evidence) vs. *epistemic* (policy) uncertainty. Filtered variants restrict to schema labels $\mathcal{A} = \{0, 1, 2, 3\}$; $\varepsilon = 10^{-12}$ ensures stability.

Feature Family	Mathematical Formula	Description / Intuition
A. Outcome Distribution: Unfiltered Top-k (Gray-Box Access)		
Context: Given top-k log-probabilities $\{\ell_1, \ldots, \ell_k\}$ for the outcome token (we use k=5), compute renormalized probabilities $\tilde{p}_i = \exp(\ell_i)/\sum_{j=1}^{k} \exp(\ell_j)$.		
Entropy (base-2)	$H_2(\tilde{p}) = -\sum_{i=1}^{k} \tilde{p}_i \log_2 \tilde{p}_i$	Uncertainty over the outcome distribution; higher values indicate greater indecision.
Normalized Entropy	$H_2^{\text{norm}}(\tilde{p}) = \frac{H_2(\tilde{p})}{\log_2 k}$	Entropy scaled to $[0, 1]$, invariant to k.
Effective Choices	$N_{\text{eff}} = 2^{H_2(\tilde{p})}$	Number of equally likely outcomes consistent with the entropy (interpretable as perplexity).
Confidence Score (Top-k)	$C_{\text{top-}k} = 1 - H_2^{\text{norm}}(\tilde{p})$	Complement of normalized entropy; ranges $[0, 1]$ with 1 = fully confident.
Max Softmax Probability (MSP)	$\text{MSP} = \max_i \tilde{p}_i$	Classic confidence baseline; lower values imply greater uncertainty.
Top-2 Margin	$\Delta p = \tilde{p}_{(1)} - \tilde{p}_{(2)}$	Absolute separation between the most and second-most probable outcomes.
Top-2 Margin (Normalized)	$\Delta p^{\text{norm}} = \frac{\tilde{p}_{(1)} - \tilde{p}_{(2)}}{\max\{\tilde{p}_{(1)}, \varepsilon\}}$	Relative margin; ε=10^{-12} prevents division by zero.
Top-1/Top-2 Ratio	$R_{1/2} = \frac{\tilde{p}_{(1)}}{\max\{\tilde{p}_{(2)}, \varepsilon\}}$	Confidence ratio between top two candidates.
B. Outcome Distribution: Filtered to Valid Schema Labels (Gray-Box Access)		
Context: Request top-20 log-probabilities, filter to schema-valid labels $\mathcal{A}$ (binary: $\{0, 1\}$; expanded: $\{0, 1, 2, 3\}$), collapse token mass onto each label, renormalize to obtain $\tilde{p}^{(\mathcal{A})}$, then recompute metrics.		
Filtered Entropy	$H_2(\tilde{p}^{(\mathcal{A})})$	Uncertainty conditional on valid labels only; robust to spurious high-probability tokens.
Filtered Normalized Entropy	$H_2^{\text{norm}}(\tilde{p}^{(\mathcal{A})})$	Filtered entropy scaled to $[0, 1]$.
Filtered Effective Choices	$2^{H_2(\tilde{p}^{(\mathcal{A})})}$	Effective number of schema-consistent outcomes.
Filtered Confidence Score	$1 - H_2^{\text{norm}}(\tilde{p}^{(\mathcal{A})})$	Confidence over valid decision classes.
Filtered Top-2 Margin	$\Delta p^{(\mathcal{A})} = \tilde{p}_{(1)}^{(\mathcal{A})} - \tilde{p}_{(2)}^{(\mathcal{A})}$	Margin between top two valid labels.
Filtered Top-2 Margin (Normalized)	$\frac{\tilde{p}_{(1)}^{(\mathcal{A})} - \tilde{p}_{(2)}^{(\mathcal{A})}}{\max\{\tilde{p}_{(1)}^{(\mathcal{A})}, \varepsilon\}}$	Normalized filtered margin.
Filtered Top-1/Top-2 Ratio	$\frac{\tilde{p}_{(1)}^{(\mathcal{A})}}{\max\{\tilde{p}_{(2)}^{(\mathcal{A})}, \varepsilon\}}$	Confidence ratio over valid labels.
C. Log-Probability Margin (Gray-Box Access)		
Log-Odds Margin (Top-2)	$\Delta_\ell = \ell_{(2)} - \ell_{(1)}$	Difference in log-space; more negative = stronger preference for top outcome.
Normalized Log-Odds Margin	$\Delta_\ell^{\text{norm}} = \frac{\ell_{(2)} - \ell_{(1)}}{\ell_{(2)}}$	Relative Margin; stabilizes comparison across models/prompts.
Filtered Log-Odds Margin	$\Delta_\ell^{(\mathcal{A})} = \ell_{(2)}^{(\mathcal{A})} - \ell_{(1)}^{(\mathcal{A})}$	Margin computed over filtered valid labels.
Filtered Normalized Top 2 Margin	$\frac{\ell_{(2)}^{(\mathcal{A})} - \ell_{(1)}^{(\mathcal{A})}}{\ell_{(2)}^{(\mathcal{A})}}$	Normalized version for filtered labels.
D. Reasoning Sequence-Level Features (Gray-Box Access, CoT Required; not used in final reported results)		
Context: Given a CoT reasoning sequence of T tokens $\{y_1, \ldots, y_T\}$ with per-token log-probabilities $\{\log p(y_t)\}_{t=1}^{T}$.		
Sequence Negative Log-Likelihood	$\text{NLL} = -\sum_{t=1}^{T} \log p(y_t)$	Total surprise of the reasoning sequence; higher = less confident generation.
Perplexity (Natural Base)	$\text{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log p(y_t)\right)$	Standard perplexity metric; higher values indicate less fluent reasoning.
E. Reasoning Token-Level Distributional Features (Gray-Box Access, CoT Required; not used in final reported results)		
Context: For each reasoning token y_t, compute per-token entropy $h_t = -\sum_j \tilde{q}_{tj} \log_2 \tilde{q}_{tj}$ where $\tilde{q}_{tj}$ are renormalized top-k probabilities. Aggregate over T tokens.		
Mean Token Entropy	$\bar{h} = \frac{1}{T}\sum_{t=1}^{T} h_t$	Average per-token uncertainty during generation.
Token Entropy Quantiles	$Q_q(\{h_t\}),\ q \in \{0, 0.25, 0.5, 0.75, 1.0\}$	Distributional shape of per-token uncertainties .
Token Probability Quantiles	$Q_q(\{p(y_t)\}),\ q \in \{0, 0.25, 0.5, 0.75, 1.0\}$	Distributional quantiles of per-token likelihoods.
F. Verbalized (Self-Reported) Confidence (Black-Box Compatible)		
Reported Confidence (Scalar)	$\hat{c} \in [0, 100]$ (normalized to $[0, 1]$)	LLM's explicit self-assessment of confidence, extracted from structured output.
Confidence Bands (One-Hot)	$\mathbb{I}\{\text{VL, L, M, H, VH}\}$	Coarse-grained verbalization: Very Low, Low, Medium, High, Very High.
G. Uncertainty Attribution Indicators (Black-Box Compatible)		
Evidence-Deficit Indicator (Binary)	$\mathbb{I}\{\text{outcome} = 2\}$	1 if the LLM abstained due to insufficient evidence (aleatoric); 0 otherwise.
Policy-Gap Indicator (Binary)	$\mathbb{I}\{\text{outcome} = 3\}$	1 if the LLM abstained due to a definition/policy gap (epistemic); 0 otherwise.

Table 7: Cost-ratio sensitivity analysis. Relative expected cost $C(\tau^\star)/c_{\mathbf{mis}}$ recomputed for cost ratios $r = c_{\mathbf{rev}}/c_{\mathbf{mis}} \in [0.4, 0.9]$, reported at representative values $\{0.4, 0.64, 0.9\}$ (range endpoints and baseline).

(a) OpenAI Moderation Dataset

LLM	$r = 0.4$	$r = 0.64$	$r = 0.9$
gpt-4o-mini	78	38	178
gpt-4o	106	51	190
gpt-4.1-mini	163	45	244
gemini-2.5-flash-lite	187	77	285
gemini-2.0-flash-lite	97	41	194
gemini-2.0-flash-001	201	69	289
QWEN3-14B	121	56	212
QWEN3-32B	NA	NA	NA
LLAMA32-11B	NA	NA	NA

(b) Multimodal Moderation Dataset

LLM	$r = 0.4$	$r = 0.64$	$r = 0.9$
gpt-4o-mini	72	22	96
gpt-4o	18	29	63
gpt-4.1-mini	20	30	65
gemini-2.5-flash-lite	0	40	54
gemini-2.0-flash-lite	54	45	100
gemini-2.0-flash-001	0	34	44
QWEN3-14B	53	36	93
QWEN3-32B	31	47	85
LLAMA32-11B	71	39	108