Token-Level Marginalization for Multi-Label LLM Classifiers

Anonymous ACL submission

Abstract

This paper addresses the critical challenge of deriving interpretable confidence scores from generative language models (LLMs) when applied to multi-label content safety classification. While models like LLaMA Guard are effective for identifying unsafe content and its categories, their generative architecture inherently lacks direct class-level probabilities, which hinders model confidence assessment and performance interpretation. This limitation complicates the setting of dynamic thresholds for content moderation and impedes fine-grained error analysis. This research proposes and evaluates three novel token-level probability estimation approaches to bridge this gap. The aim is to enhance model interpretability and accuracy, and evaluate the generalizability of this framework across different instruction-tuned models. Through extensive experimentation on a synthetically generated, rigorously annotated dataset, it is demonstrated that leveraging token logits significantly improves the interpretability and reliability of generative classifiers, enabling more nuanced content safety moderation.

1 Introduction

The rise of user-generated content has heightened the importance of content safety on digital platforms. Effective moderation systems must not only detect harmful content but also accurately categorize violations. Large Language Models (LLMs), known for their robust language understanding, are increasingly central to this task (Padhi et al., 2024; Zeng et al., 2024; Inan et al., 2023). Models like LLaMA Guard (Inan et al., 2023) have been adapted for multi-label classification, producing structured outputs such as `'unsafe\nS1, S3'`, aligned with a predefined safety taxonomy.

However, generative models like LLaMA Guard lack native support for producing confidence scores per predicted label, unlike discriminative classifiers. This absence complicates tasks such as thresholding, prioritization, and error analysis, which are critical in high-stakes settings (Geng et al., 2024; Detommaso et al., 2024; Tian et al., 2023). Without interpretable confidence, such systems risk both over-censorship and under-moderation.

To mitigate this, we introduce a framework that derives category-level confidence scores from token-level probabilities during autoregressive decoding. Following prior work (Cheng et al., 2024; Zhang et al., 2025), we evaluate three types of uncertainty estimation strategies: conditional probability, joint probability, and marginal probability.

Contributions:

1. A principled method to extract confidence scores from generative LLMs;
2. A comparison of multiple probability estimation techniques;
3. Demonstration of the method's generalizability to instruction-tuned models.

2 Related Work

Recent work has explored deriving confidence estimates from generative large language models (LLMs) (Ma et al., 2025; Xia et al., 2025; Yang et al., 2024; Vashurin et al., 2025). Given their token-by-token decoding mechanism, researchers have proposed using logits and log-probabilities to estimate uncertainty (Mena et al., 2021; Vazhentsev et al., 2025; Yang et al., 2025). Log-probabilities which are obtained via softmax over logits, can provide token or sequence-level likelihoods through methods such as joint or conditional probability aggregation (Fadeeva et al., 2024).

Methods like Logits-induced Token Uncertainty (LogU) compute token-level uncertainty efficiently without sampling, enabling applications in reranking and prompt engineering (Ma et al., 2025).

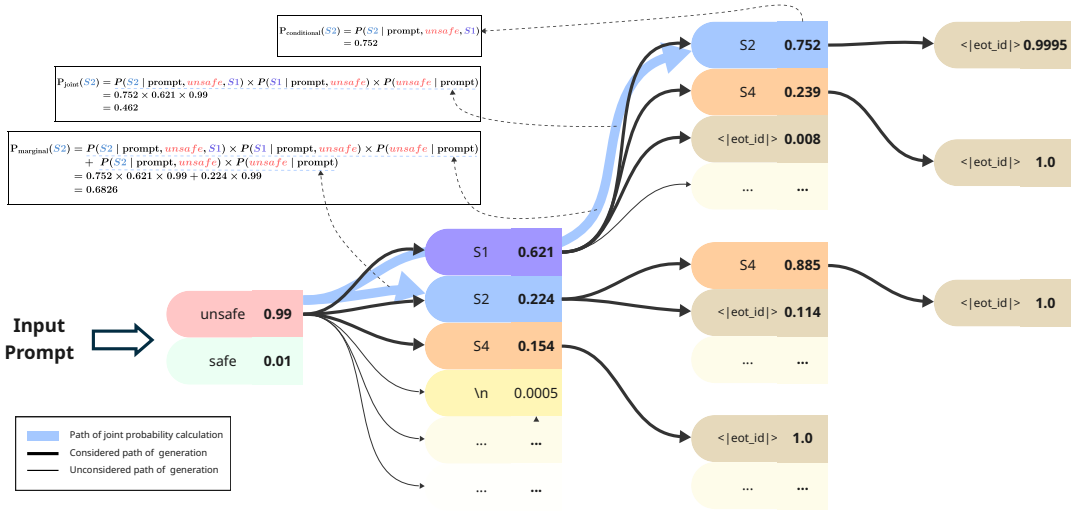

Figure 1: We explore Conditional, Joint, and Marginal probability-based approaches to estimate model confidence. The category labels (e.g., S1, S3, etc.) correspond to classes defined in the LLaMA Guard taxonomy and are treated as tokens for simplicity.

Claim-conditioned probability estimation has been used to assess uncertainty around specific factual claims in tasks such as fact-checking (Fadeeva et al., 2024), and prompt recovery techniques like LOGIT2PROMPT leverage similar signal (Morris et al., 2023).

However, most prior work focuses on single-label classification or holistic sequence scoring. The challenge of systematically mapping token-level uncertainty to structured multi-label confidence remains underexplored.

3 Methodology

3.1 Problem Formulation: Generative Models as Multi-label Classifiers

Formally, let X represent a textual content instance (input) and let Y denote the set of K predefined safety categories, $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$. For a multi-label classification task, an input instance x can be associated with any subset of these labels, i.e., $y \subseteq \mathcal{C}$. This can be represented by a binary vector $y = [y_1, y_2, \ldots, y_K]$, where $y_i = 1$ if category C_i is violated, and $y_i = 0$ otherwise.

Generative LLMs, such as LLaMA Guard, are trained to model the joint probability distribution of input and output tokens, $P(X, T)$, or, more commonly, the conditional probability of output tokens given the input, $P(T \mid X)$. Here,

$$T = (t_1, t_2, \ldots, t_L)$$

represents the generated sequence of tokens that constitutes the classification output (e.g., `"unsafe\nS1, S3"`).

The fundamental challenge lies in deriving interpretable and reliable category-level confidence scores, $P(y_i = 1 \mid X)$, from this generative output, which is a sequence of tokens rather than explicit class probabilities.

3.2 Token-Level Probability Estimation Approaches

All proposed methods leverage the raw, unnormalized scores (logits) generated by the LLM's final layer for each token in its vocabulary. These logits are then transformed into probabilities via a softmax function.

3.2.1 Conditional Probability

This approach computes the likelihood of a label token (e.g., `"S1"`) appearing at a specific step in the output, conditioned on the input prompt and previously generated tokens. It reflects the model's immediate probability of generating a given safety label during decoding.

For a target label C_i represented by token(s) t_{C_i}, its conditional probability at generation step j is:

$$P(t_j = t_{C_i} \mid X, t_1, \ldots, t_{j-1}), \quad (1)$$

which is obtained directly from the softmax output at step j.

In multi-label settings, we identify the label tokens (e.g., `"S1"`, `"S3"`) in the output and log their

probabilities at generation. [1]

3.2.2 Joint Probability

This method computes the joint probability of generating each individual token in the output, conditioned on the input prompt and all previously generated tokens. For any target token t_j in the generated sequence $T = (t_1, t_2, \ldots, t_L)$, the joint probability up to and including t_j is given by:

$$P(t_{\leq j} \mid X) = P(t_1 \mid X) \times P(t_2 \mid X, t_1) \times \cdots \times P(t_j \mid X, t_1, \ldots, t_{j-1}) \quad (2)$$

In practice, to improve numerical stability, the logarithm of the joint probability is computed as a sum of log probabilities.

3.2.3 Marginal Probability

Marginal probability estimates the overall likelihood of a specific label C_i appearing in the model's output, considering all possible sequences containing that label, given an input X:

$$P(C_i \mid X) = \sum_{T \in \mathcal{T}_{C_i}} P(T \mid X), \quad (3)$$

where $\mathcal{T}_{C_i}$ denotes the set of output sequences that include C_i. The joint probability of each sequence $T = (t_1, \ldots, t_L)$ is given by:

$$P(T \mid X) = \prod_{j=1}^{L} P(t_j \mid X, t_{<j}). \quad (4)$$

While theoretically comprehensive, capturing the true likelihood of label presence, this formulation is computationally intractable due to the exponential size of $\mathcal{T}_{C_i}$.

To approximate this, we adopt a constrained decoding strategy, detailed in Appendix A.1

1. **Top-p Filtering**: At each step, only tokens whose cumulative probability is below a threshold (e.g., 0.99) are considered, following nucleus sampling to prune unlikely paths.

2. **Maximum generation depth**: We set a limit on the maximum number of tokens that can be generated along any given path.

3. **Early Stopping on `[EOS]`**: Decoding halts upon generating an end-of-sequence token, ensuring that only complete outputs contribute to the final estimate.

This approximation balances tractability with fidelity, enabling category-level marginal probability estimation in practice.

The distinction between conditional, joint, and marginal probabilities is crucial, as each offers a unique perspective on the model's confidence. Conditional probability focuses on the likelihood of individual label tokens at their point of generation. Joint probability assesses the confidence of the entire predicted label string. Marginal probability, being the most complex, attempts to capture the overall likelihood of a label independent of its exact position or co-occurrence with other specific tokens in the output string.

3.3 Data Generation and Annotation

As LLaMA Guard has not officially released any test datasets and no publicly available benchmark exists that aligns with its taxonomy, we opted to synthetically generate the evaluation data (Appendix A.2). Each content instance in the synthetic dataset is crafted to violate at least 2–3 safety categories based on LLaMA Guard 3 taxonomy. This controlled generation ensures a diverse set of multi-label examples, allowing for comprehensive evaluation across various safety categories.

To ensure accurate ground truth labels, each data point's category annotations are derived by three separate LLMs. Only examples with at least 2 out of 3 model agreements matching the ground truth are retained. This reconciliation strategy creates a highly reliable "gold standard" dataset for evaluation. The final evaluation dataset consists of 2.3k records, with each category containing between 229 and 491 samples.

4 Evaluation

4.1 Benchmarks

We evaluate our approaches using greedy decoding across all models. For comparison, we include uncertainty estimation techniques introduced by (Ma et al., 2025), namely Probability Uncertainty, Entropy Uncertainty, and LogTokU. In addition to our synthetically generated dataset, we incorporate the Beavertails benchmark (Ji et al., 2023) to assess the performance of different methods under standardized evaluation settings.

[1] When labels span multiple tokens (as in LLaMA Guard, where `"S1"` is tokenized as `'S'`, `'1'`), the probability of the final token (e.g., `'1'`) is used as a proxy for the label's likelihood.

Model	Method	Synthetic Dataset		Beavertails	
		F1	AUCROC	F1	AUCROC
LLaMA Guard 2	Greedy Generation	0.644	–	0.430	–
	Probability Uncertainty	0.496	–	0.431	–
	Entropy Uncertainty	0.496	–	0.431	–
	LogTokU	0.533	–	0.430	–
	Conditional Probability	0.644	0.756	0.430	0.649
	Joint Probability	0.644	0.754	0.430	0.649
	Marginal Probability	**0.658**	**0.824**	**0.442**	**0.705**
LLaMA Guard 3	Greedy Generation	0.701	–	0.418	–
	Probability Uncertainty	0.698	–	0.417	–
	Entropy Uncertainty	0.698	–	0.417	–
	LogTokU	0.669	–	0.418	–
	Conditional Probability	0.701	0.777	0.418	0.640
	Joint Probability	0.700	0.777	0.419	0.640
	Marginal Probability	**0.768**	**0.906**	**0.449**	**0.805**
LLaMA 3.1 8B Instruct	Greedy Generation	0.697	–	0.373	–
	Probability Uncertainty	0.453	–	0.420	–
	Entropy Uncertainty	0.453	–	0.420	–
	LogTokU	0.462	–	**0.426**	–
	Conditional Probability	0.704	0.876	0.376	0.677
	Joint Probability	0.626	0.874	0.401	0.678
	Marginal Probability	**0.738**	**0.934**	0.424	**0.809**

Table 1: Comparison of various methods across multiple LLM-based safety classifiers. ↑ indicates higher-is-better metrics; ↓ indicates lower-is-better.

4.2 Evaluation Metrics

We evaluate model performance using standard metrics for multi-label classification: F1-score and AUCROC. F1-score is the harmonic mean of precision and recall. We report micro-averaged F1 across all labels to capture overall performance. AUCROC evaluates the model's ability to distinguish between positive and negative classes. For multi-label settings, it is averaged over all labels.

4.3 Results

The primary models considered for content safety classification is the LLaMA Guard models. Its direct classification output (e.g., "unsafe
nS1, S3") will serve as the baseline for performance comparison.

The evaluation of the Conditional, Joint, and Marginal probability methods (outlined in Section 3.2) for deriving category-level confidence scores on the LLaMA Guard model is shown in Table 1. The results show that leveraging token logits for probability estimation significantly improves classification performance. The results shows that the Marginal Probability, leveraging its ability to aggregate probabilities across multiple paths, provides the most robust and accurate confidence scores, leading to superior overall classification performance.

4.4 Generalizability of the approach

To assess the transferability of the proposed strategy, LLaMA 3.1-8B-Instruct is considered. This model is an instruction-tuned LLM that has not been explicitly fine-tuned for content safety. The proposed probability-based decoding approach will be applied to the model and the performance of our approach is compared against the vanilla greedy decoding approach. Evaluation results in Table 1 show that even without explicit safety fine-tuning, a general instruction-tuned model, when used in the multi-label classification setting would show improved performance with marginal probability based approach.

5 Conclusion

This paper addressed the challenge of deriving interpretable confidence scores from generative LLMs for multi-label content safety classification. We proposed three token-level probability estimation methods—Conditional, Joint, and Marginal—to extract confidence scores from token logits. Experiments on a synthetic dataset show that these methods, especially the Marginal approach, significantly enhance classification accuracy. Overall, this work demonstrates that generative models can be adapted into reliable, interpretable multi-label classifiers, enabling broader use.

5.1 Limitations and Future Work

This work presents a novel approach for deriving confidence scores from generative LLMs in multi-label settings, but several limitations remain.

First, evaluations were performed on synthetic datasets. While useful for controlled experimentation, such data may not fully reflect the complexity and ambiguity of real-world harmful content, despite efforts to simulate realistic label distributions.

Second, the marginal probability estimation is approximate and does not explore the full space of generation paths. While tractable, this limits accuracy. Future work could investigate more efficient or principled marginal estimation techniques and examine how decoding strategies (e.g., beam width, top-p sampling) affect robustness.

Finally, the marginal probability method incurs token-level overhead due to multiple path explorations, which may hinder real-time applications. Practical deployment will require strategies to reduce this cost, such as adaptive path selection or approximation schemes.

References

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. *arXiv preprint arXiv:2406.14491*.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. 2024. Multicalibration for confidence scoring in llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

dphn. 2024. dolphin-2.9.2-phi-3-medium-abliterated. https://huggingface.co/dphn/dolphin-2.9.2-Phi-3-Medium-abliterated. Accessed: 2025-07-29.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

huihui-ai. 2024a. Llama-3.3-70b-instruct-abliterated. https://huggingface.co/huihui-ai/Llama-3.3-70B-Instruct-abliterated. Accessed: 2025-07-29.

huihui-ai. 2024b. Qwq-32b-abliterated. https://huggingface.co/huihui-ai/QwQ-32B-abliterated. Accessed: 2025-07-29.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. 2025. Estimating llm uncertainty with evidence. *arXiv preprint arXiv:2502.00290*.

José Mena, Oriol Pujol, and Jordi Vitrià. 2021. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35.

John X Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. 2023. Language model inversion. *arXiv preprint arXiv:2311.13647*.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, and 1 others. 2024. Granite guardian. *arXiv preprint arXiv:2412.07724*.

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. *arXiv preprint arXiv:2305.14975*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025. Token-level density-based uncertainty quantification methods for

eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. *arXiv preprint arXiv:2503.00172*.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, and 1 others. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, and 1 others. 2025. Token-level uncertainty estimation for large language model reasoning. *arXiv preprint arXiv:2505.11737*.

A Appendix

A.1 Marginal Probability Algorithm

Algorithm 1: Compute Marginal Probability of Label via Beam-like DFS with Max Token Cutoff

Function `ComputeMarginalProbability(`*inputs, labels, max_new_tokens*`):`
  Initialize probabilities[label] $\leftarrow 0$ for each label
  Procedure `DFS(`*inputs, current_probability, depth*`):`
    If *current_probability* $< 1e^{-7}$ **Return Generate** next token logits using model
    top_tokens $\leftarrow$ Get top-p tokens with their probabilities
    For each *(token, probability) in top_tokens* new_inputs $\leftarrow$ Append token to input_ids and attention_mask
    generation $\leftarrow$ Decode new_inputs to text
    For each *label in labels* **If** *generation ends with label* probabilities[label] $+= current\_probability \times probability$
    If *token is EOS and probability* ≥ 0.7 **break** `// Stop exploring this path`
    If *EOS token is among top tokens and this is the third token* **break** `// Stop exploring this path`
    If *depth = max_new_tokens or token is EOS* **continue** `// Skip recursion`
    Call `DFS(`*new_inputs, current_probability* $\times$ *probability, depth + 1*`)`
  Call `DFS(`*inputs, 1.0, 0*`)`
  Return probabilities

A.2 Synthetic Data Generation

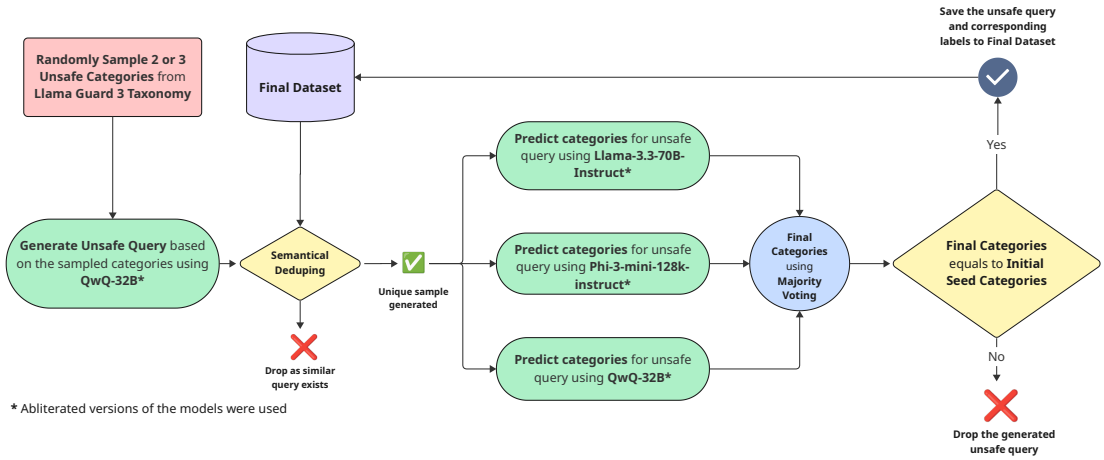

Figure 2: An overview of the synthetic data generation pipeline used for generating the evaluation data. The models employed in this process include Qwen/QwQ-32B (huihui-ai, 2024b), Meta-Llama/Llama-3.3-70B-Instruct (huihui-ai, 2024a), and Microsoft/Phi-3-mini-128k-Instruct (dphn, 2024). Abliterated versions of these models were utilized to enable the generation of unsafe and offensive content.