FROM SMALL TO LARGE LANGUAGE MODELS: HOW MUCH CONFIDENCE CAN WE HAVE?

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we provide novel insights into information-based, consistency-based and self-verbalized uncertainty quantification (UQ) for multi-label text classification across a range of recent language models on a new unsaturated benchmark of medical device adverse event reports with interdependent labels. We compare more than twenty encoder- and decoder-only language models across three paradigms: discriminative fine-tuning, generative fine-tuning, and few-shot in-context prompting (instruction-tuned and reasoning variants, local and API-accessible). UQ is performed using token-information measures, consistency under stochastic generation, and self-verbalized confidence, with utility assessed via selective prediction. We provide practical guidance on model selection, when fine-tuning is preferable to prompting, and which UQ signals are most effective for routing and human-in-the-loop triage. Our results reveal trade-offs across model types. Discriminatively fine-tuned decoders achieve the strongest head–tail accuracy while still offering solid uncertainty quantification (UQ). In contrast, generative fine-tuning provides the most reliable UQ overall. Reasoning models improve performance on extreme-tail labels but yield weak UQ. Finally, self-verbalized confidence proves unreliable as an indicator of model certainty.

1 INTRODUCTION

Large language models (LLMs) are increasingly being applied in domains such as education, academia, and industry. While their use is highly efficient and generally accessible to a broad audience, the development of these systems remains concentrated within a relatively small group of organizations. This imbalance has sparked interest in smaller language models, which offer users and developers more flexibility to adapt performance and to explore additional aspects of trustworthiness, such as uncertainty quantification (UQ). Although both small and large language models have already demonstrated remarkable levels of performance, accuracy alone is often not sufficient to ensure broad adoption. Trust-related criteria are equally important, especially in high-stakes contexts like healthcare, where the need for human oversight becomes paramount (Shneiderman, 2020). Nevertheless, despite their relevance, considerations of trustworthiness are still widely underexplored and often overlooked in the existing literature. Many of the aforementioned aspects depend further on the specific use case. Multi-label classification, though highly relevant in practice, remains both understudied and often overlooked in the literature. It is central to a wide range of applications, including document categorization, clinical coding, and email filtering. Language models could play an important role in tackling such tasks, provided they can be applied in a trustworthy manner. However, the rapid pace of model development and availability presents practitioners with challenging choices: which model (e.g., ModernBERT, LLaMA, etc.) is most suitable? If necessary, which learning paradigm should be adopted (e.g., fine-tuning, in-context learning, etc.)? And how reliable are the resulting predictions? Effective UQ can provide a pathway toward more reliable and trustworthy systems (Ojha et al., 2025; Liu et al., 2024). However, as methods differ substantially between discriminative and generative language models, it is essential to understand and select the appropriate approach for each model and learning paradigm. For smaller, discriminative models, entropy-based measures of uncertainty are relatively well studied and better understood (Baur et al., 2025). In contrast, for large generative models, uncertainty quantification becomes more challenging and complex: token-level uncertainty, consistency across multiple generations, and self-verbalized confidence all introduce new difficulties and might be miscalibrated. As a result, determining the

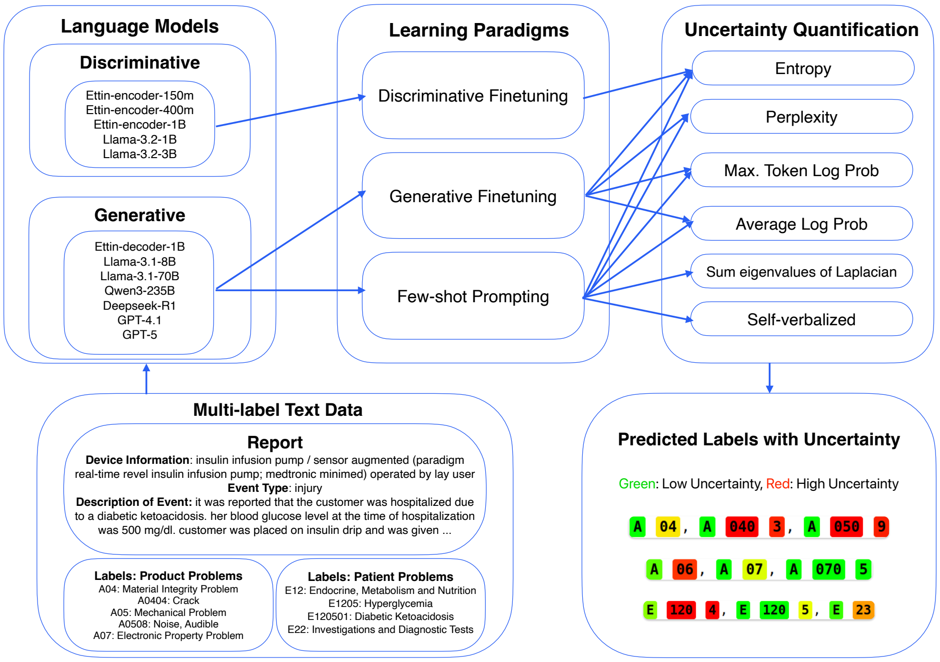
Language Models
Discriminative
Ettin-encoder-150m
Ettin-encoder-400m
Ettin-encoder-1B
Llama-3.2-1B
Llama-3.2-3B
Generative
Ettin-decoder-1B
Llama-3.1-8B
Llama-3.1-70B
Qwen3-235B
Deepseek-R1
GPT-4.1
GPT-5
Learning Paradigms
Discriminative Finetuning
Generative Finetuning
Few-shot Prompting
Uncertainty Quantification
Entropy
Perplexity
Max. Token Log Prob
Average Log Prob
Sum eigenvalues of Laplacian
Self-verbalized
Multi-label Text Data
Report
Device Information: insulin infusion pump / sensor augmented (paradigm real-time revel insulin infusion pump; medtronic minimed) operated by lay user
Event Type: injury
Description of Event: it was reported that the customer was hospitalized due to a diabetic ketoacidosis. her blood glucose level at the time of hospitalization was 500 mg/dl. customer was placed on insulin drip and was given ...
Labels: Product Problems
A04: Material Integrity Problem
A0404: Crack
A05: Mechanical Problem
A0508: Noise, Audible
A07: Electronic Property Problem
Labels: Patient Problems
E12: Endocrine, Metabolism and Nutrition
E1205: Hyperglycemia
E120501: Diabetic Ketoacidosis
E22: Investigations and Diagnostic Tests
Predicted Labels with Uncertainty
Green: Low Uncertainty, Red: High Uncertainty
A 04 , A 040 3 , A 050 9
A 06 , A 07 , A 070 5
E 120 4 , E 120 5 , E 23

Figure 1: Schematic overview of the experimental setup, with an exemplary text report, hierarchical labels (product/patient problems; hierarchy indicated by the digits; e.g., label 'A0404' is a child of 'A04'), models and learning paradigms, uncertainty quantification methods, and predicted labels with (color coded) uncertainty. We finally assess the predictive performance and quality of the uncertainty quantification (not depicted).

most effective UQ strategy – for purposes such as sample routing and efficient human-in-the-loop triage – remains an open research question.

Our contribution. To avoid reliance on saturated benchmarks and potential pre-training exposure, we constructed a new dataset of medical device event reports, serving as a prototypical use case for exploring multi-label classification not only in terms of predictive performance but also across various aspects of uncertainty quantification (Figure 1), offering insights into trustworthiness and reliability across over 20 popular language models and different UQ methods. In particular, we (1) create a new unsaturated multi-label text classification dataset with a skewed distribution of hierarchical and thus interdependent labels; (2) fine-tune over twenty encoder- and decoder-only models in discriminative and generative settings, and compare them with zero-/few-shot prompting of locally-hosted decoders and API-accessible models with the proposed dataset; (3) systematically study multiple UQ approaches tailored to these settings – including information-, consistency- and self-verbalized based uncertainty – and study their quality (and thus ultimately, their utility for sample routing or triage).

Overall, our results provide practical guidance on language model choice, on when fine-tuning is preferable to in-context learning, and on how to select and calibrate UQ methods, such that they are suited for routing ambiguous samples to stronger but more expensive models or flagging them for secondary human review.

2 Related Work

2.1 Multi-label text classification

Early approaches to multi-label text classification relied on classical models such as Naïve Bayes, support vector machines (Wang & Manning, 2012), or recurrent neural networks with mixed results.

With the advent of transformer-based language models (Vaswani et al., 2017), both encoder-only (Huang et al., 2021) and decoder-only (Ma et al., 2025; Galke et al., 2025) architectures have been applied to this task. Encoder-only models are typically equipped with a linear classification head and discriminatively fine-tuned on multi-labeled text, usually with a variant of the binary cross-entropy loss. Decoder-only models are often fine-tuned generatively—by producing label tokens via a softmax decoder head — and can also be adapted for discriminative training with a classification head; however, controlled comparisons between generative and discriminative fine-tuning for decoder models remain scarce. Head-to-head comparisons between encoder- and decoder-only models under matched conditions are also rare, largely because publicly available suites have historically lacked paired models with comparable parameter counts and training data. Weller et al. (2025) address this by releasing a suite of matched encoder-only and decoder-only models (‘Ettin’) trained on the ModernBERT architecture (Warner et al., 2024b) across model sizes (from 17m parameters to 1b parameters), but they do not evaluate on multi-label classification, leaving this comparison open.

Decoder-only LLMs support zero-shot prompting – performing inference without using any labeled data – and few-shot prompting via in-context learning, i.e., conditioning on a small set of (random or carefully selected) labeled exemplars without updating model weights. This has been compared with fine-tuning on a variety of tasks where medical expertise is required, with differing results. Nori et al. (2023) and Maharjan et al. (2024) found that in-context learning outperforms fine-tuning on question-answering. In contrast, Labrak et al. (2024) found that fine-tuning surpassed few-shot prompting on question-answering. Across broader benchmarks, the relative advantage of fine-tuning versus prompting varies by task, data regime, and model (Chen et al., 2025; Wu et al., 2025). For multi-label classification specifically, performance is highly sensitive to prompt design, label semantics, output constraints, and thresholding strategies, motivating controlled comparisons.

Public benchmarks for multi-label text classification span diverse domains, including newswire (‘Reuters-21578’, Lewis, 1997), scientific literature (e.g., ‘Arxiv Academic Paper Dataset’, Yang et al., 2018; ‘BioCreative LitCovid’, Chen et al., 2022; ‘ACL Anthology’, Schopf et al., 2023; ‘MFHAD2’, Fallah et al., 2022), finance-related user posts (‘Money.StackExchange’, Maia et al., 2021), patents (‘Patent_CIRCA_45k’, Tang et al., 2020)), clinical notes for ICD coding (‘MIMIC-III’, Johnson et al., 2016; Mullenbach et al., 2018), legislative documents (‘EuroVoc’, Steinberger et al., 2012, Boella et al., 2013; ‘EURLEX57K’, Chalkidis et al., 2019; ‘KEVLAR’, Bocchi et al., 2024), and more. Most of these are scraped from online repositories and rely on distant or weak supervision (e.g., tags/metadata as labels), often with limited documentation, which introduces label noise and hampers reproducibility. Multiple sources, preprocessing choices, copyright limitations, and dataset versions further complicate fair comparisons. Heavy reuse has also lead to benchmark saturation via overfitting and implicit adaptation. In addition, overlap between benchmark content and LLM pre-training corpora raises contamination concerns that can inflate zero-/few-shot performance. These challenges motivate periodically introducing better-documented, decontaminated benchmarks with standardized splits to assess generalization.

2.2 Uncertainty quantification

For quantifying uncertainty in machine learning models, numerous approaches have been proposed. Confidence calibration for imbalanced data has been reviewed by Dong et al. (2025). Token- or information-based uncertainty metrics, which in various forms rely on the log-probabilities of an LLM’s outputs, have been surveyed and formalized by Shorinwa et al. (2025); Fomicheva et al. (2020). Among those of interest are: entropy of the top-n log-probabilities, perplexity, maximum token log-probability, and the average token log-probability. Consistency-based metrics, which estimate uncertainty by measuring the variability of a model’s outputs under input perturbations, prompt variations, or model stochasticity, have been widely used for LLMs as a way to capture epistemic uncertainty Xiao et al. (2025). Lin et al. (2023) suggested an effective approach for deriving a consistency-based uncertainty score through the sum of eigenvalues of a Laplacian graph. Combining information-based and consistency-based metrics has been proposed by Vashurin et al. (2025) as an effective strategy to improve uncertainty estimation, motivating our approach of multiplicatively integrating token-level uncertainty with consistency measures. Finally, self-verbalized uncertainty, in which a model outputs a confidence score along with its prediction, has been explored by Tian et al. (2023) and Harsha Tanneru et al. (2024), offering a complementary approach to directly capture model-reported uncertainty, especially in cases where log-probabilities are not available.

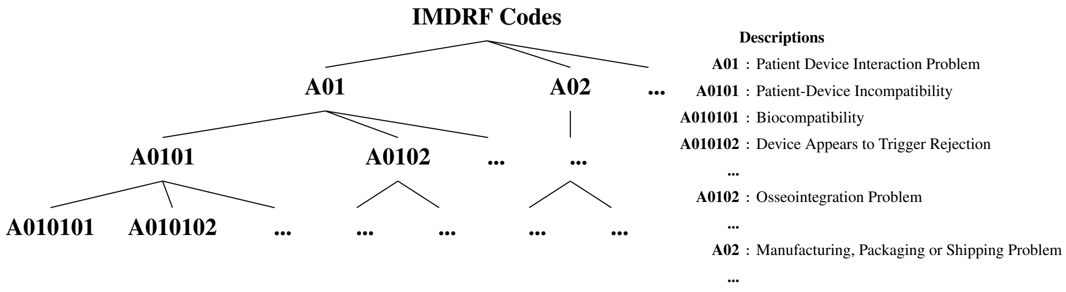
IMDRF Codes
A01
A02
...
A0101
A0102
...
...
A010101
A010102
...
...
...
...
...
Descriptions
A01 : Patient Device Interaction Problem
A0101 : Patient-Device Incompatibility
A010101 : Biocompatibility
A010102 : Device Appears to Trigger Rejection
...
A0102 : Osseointegration Problem
...
A02 : Manufacturing, Packaging or Shipping Problem
...

Figure 2: IMDRF codes refer to product (device) and patient problems (defined in the different annexes A, B, C, etc. of the IMDRF terminology) and comprise the hierarchical labels of the dataset. IMDRF code 'A010101' is a child of 'A0101' and grandchild of 'A01' in the code hierarchy (a graph not a tree).

3 Data – FDA Medical Device Adverse Event Reports

The US Food and Drug Administration (FDA) publishes quarterly reports of medical device adverse events at `https://open.fda.gov` alongside expertly annotated codes capturing the associated problems. We used this to curate a large-scale text classification dataset comprising approximately 5.5 million medical device adverse event reports, each annotated with hierarchical multi-labels, cf. Figure 2. Each report contains a narrative description of the event, metadata, and human-assigned labels that detail relevant product and patient problems.

3.1 Dataset creation

We collect quarterly published (zipped JSON) files spanning 2015 through mid-2025, excluding reports before 2015 due to the absence of necessary product problem labels. Data from 2015–2023 are used for training, while test data comprises reports from July 2024 to June 2025, cf. Figure A.1 for a small overview of the topics addressed in the reports. Reports from the first half of 2024 form the validation split. For each report, we concatenate the narrative 'description of event or problem', harmonize all letter cases, and extract relevant metadata (including event type – for example, malfunction, injury or death – and device information such as generic name, brand name, manufacturer, and operator). Each sample is assigned a unique identifier and a timestamp denoting the report date.

3.2 Dataset labels

Labels are mapped from FDA product and patient problem terms to the International Medical Device Regulators terminology (IMDRF, 2025), which provides hierarchical codes for each product and patient problem (see Figure 2). Samples missing IMDRF codes for one or more terms are excluded. IMDRF adoption is nearly complete by 2021, but between 2015 and 2020 only about 16–20% of the samples can be assigned exclusively to IMDRF codes and are retained. To leverage the IMDRF hierarchy, we upward-propagate each code to include all ancestor codes, as determined from official IMDRF annexes. This approach recognizes that FDA annotators typically use the most specific (leaf) codes and ensures that models are not penalized for predicting valid parent codes.

Although the IMDRF label hierarchy is often tree-structured (e.g., 'A010101' is a child of 'A0101' and grandchild of 'A01'), it is technically a graph due to occasional cross-links (e.g., 'E172001' is a descendant in multiple branches). Label mapping and upward propagation are performed separately for product and patient problems, yielding two sets of hierarchical labels per report. We flatten the resulting code sets and treat the union of upward-propagated IMDRF codes as target labels. Generative models can produce these sets directly, while discriminative models use a binary multi-label vector over the IMDRF codes present in the train split.

3.3 Dataset pre-processing

IMDRF revises its label taxonomy annually, adding new terms and retiring others. To avoid introducing unseen labels into the test set, we freeze the label taxonomy as of December 2023 and discard

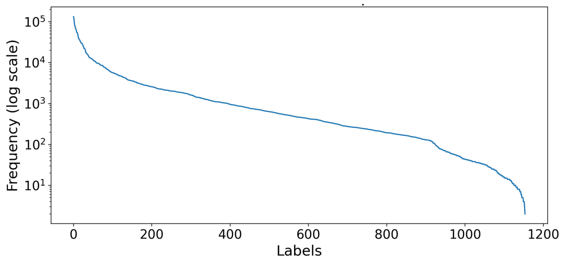

Figure 3: Distribution of label frequencies (in the combined train, validation, and truncated test set) on a logarithmic scale, showing that some labels occur very often in the dataset, while others are extremely rare.

labels introduced thereafter. We also remove extremely rare labels (fewer than 5 instances). These steps yield a final label set of 1154 labels. To reduce redundancy, we deduplicate reports based on the event description, retaining only the first occurrence, and thus reducing the dataset to 5.5 million IMDRF-compliant samples (46.4% of the original 11.9 million).

We further downsample using HDBSCAN (McInnes et al., 2017) on event descriptions, selecting cluster representatives and retaining a high proportion of rarer labels; all events labeled death are included, yielding a dataset of 488,273 samples (train: 298,825; validation: 71,271; test: 118,177). To limit inference costs, we further downsample the test set to a representative subset of 10288 samples. This 'truncated test set' set contains all of the 1154 labels at least 5 times. After tokenising with the 'cl100k_base' tokenizer (OpenAI, 2022), samples in the truncated test set average around 370 tokens with an average of 8.79 labels. On all splits, labels follow a long tail distribution; certain labels occur frequently, while others are rare (see Figure 3). In summary, our pre-processing pipeline transforms raw FDA device event data into a IMDRF terminology compliant, hierarchically labeled, deduplicated dataset tailored for a challenging multi-label classification task.

4 Experimental setup

4.1 Fine-tuning and in-context learning

Discriminative training. We fine-tune the recent Ettin suite of models (150M, 400M, 1B; Weller et al., 2025), which are based on the ModernBERT (Warner et al., 2024a) architecture. We also fine-tune Llama 3.2-1B, 3.2-3B, and 3.1-8B (Grattafiori et al., 2024) with a classification head in the final layer. All models are trained with a context length of 512 for 20 epochs each and hyperparameters (mainly learning rate and max. gradient norm.), are tuned on the validation set

Generative training. We fine-tune Llama 3.2-1B, 3.2-3B, 3.1-8B and 3.1-70B and Ettin decoder variants (400M, 1B), where the model is trained to generate class labels as tokens. Here, all models are trained for 4 epochs. An improvement for a larger number of epochs was not observed. We compare full fine-tuning with LoRA (Hu et al., 2022) for selected models.

Detailed hyperparameters for training are listed in the appendix section A.3.

Few-shot prompting and in-context learning. We host Llama 3.2-3B, 3.1-8B, and 3.1-70B, DeepSeek-R1 (37B active, 671B total; DeepSeek-AI et al., 2025), Qwen3 (4B, 30B, 235B; Yang et al., 2025), Kimi K2 (32B activated; 1T total; Kimi-Team et al., 2025), gpt-oss-120b (5.1B active, 117B total; OpenAI et al., 2025), GLM-4.5-Air (12B active, 106B total; GLM-4.5-Team et al., 2025), and Llama-3.3-Nemotron-49B-v1.5 (Bercovich et al., 2025) locally. GPT-4.1 (OpenAI, 2025b) and GPT-5 (OpenAI, 2025a) are accessed via API. Among API-based models, only GPT-4.1 provides – unlike GPT-5 – log probabilities, which are required for our UQ investigations. GPT-5 is run with medium reasoning effort and does not allow modifying temperature.

The prompt comprises the test sample, the task instructions including a list of all 1154 possible labels, and the ten nearest labeled training examples in embedding space to the inference sample. For embedding the texts, we use a pre-trained text embedding model ('bioclinical-modernbert-base-embeddings') (NeuML, 2025). Each retrieved example was truncated to 10000 characters (around 2300 tokens). Unless otherwise stated, all outputs are generated in a greedy decoding setup ($temperature = 0, top\_p = 1, top\_k = -1$). Prompts are available in the appendix A.5.

The knowledge cutoff dates for all the models are discussed in Section A.2 in the appendix.

4.2 Methods for uncertainty quantification

Uncertainty quantification for generative models. To quantify uncertainty U, we derive two complementary measures: an information-based uncertainty score U_{info} and a consistency-based uncertainty score U_{cons}. Following Vashurin et al. (2025), we combine these scores by multiplication to obtain the total uncertainty score: $U_{\text{total}} = U_{\text{info}} \cdot U_{\text{cons}}$. Here, U_{info} is computed using token-based uncertainty metrics such as entropy, average log-probability, perplexity, maximum log-probability, and response improbability (described below), while U_{cons} is computed using the Laplacian of a graph derived from multiple stochastic forward passes (Lin et al., 2024). For large language models (LLMs), we evaluate these token- and information-based uncertainty scores that serve as U_{info} :

- Entropy of Token Distribution: $H_{ij} = -\sum_{w \in D} \pi_{ij}(w) \log \pi_{ij}(w)$
- Average Token Log-Probability: $\text{Avg}(\pi) = -\frac{1}{L_i} \sum_j \log(\pi_{ij})$
- Perplexity: $\text{Perplexity}(\pi) = \exp\left(-\frac{1}{L_i} \sum_j \log(\pi_{ij})\right)$
- Maximum Token Log-Probability: $\text{Max}(\pi) = \max_j \left(-\log(\pi_{ij})\right)$
- Response Improbability: $\text{Improbability} = 1 - \prod_j \pi_{ij}$

The consistency-based score is computed by performing $n = 5$ stochastic forward passes with temperature $t = 1$, calculating the pairwise Jaccard similarity W between predictions, and computing the normalized Laplacian L as $L = I - D^{-1/2} W D^{-1/2}$, where I is the identity matrix and D is the degree matrix ($D = \text{diag}\left(\sum_j W_{ij}\right)$).

The consistency score is then $U_{\text{cons}} = \sum_k \max(0, 1 - \lambda_k)$, with λ_k denoting the eigenvalues of L.

We also evaluate self-verbalized uncertainty, for which we prompt the model to provide a confidence score $C_i \in [0, 1]$ for each prediction. We then define the self-verbalized uncertainty score as $U_{\text{self}} = 1 - C_i$, which can also serve as an independent total uncertainty score for the prediction. This formulation ensures that high confidence corresponds to low uncertainty, and vice versa.

Uncertainty quantification for discriminative models. For discriminative models in a multi-label setting, uncertainty is derived from the model's output probability vector. Since each label is binary and independent of the others, we first compute the binary entropy for each predicted label:

$$\mathbb{H}(\pi_c) = -\big[\pi_c \log(\pi_c) + (1 - \pi_c) \log(1 - \pi_c)\big],$$

where π_c is the predicted probability for the positive class (label 1) of class c. The information-based uncertainty for a sample is then the vector of these entropies across all labels:

$$U_{\text{info}} = \big(\mathbb{H}(\pi_1), \mathbb{H}(\pi_2), \ldots, \mathbb{H}(\pi_C)\big).$$

4.3 Evaluation of uncertainty scores

To evaluate the quality of uncertainty scores (either U_{total} or U_{info}), we perform selective prediction. In this setup, a model's outputs are rejected based on their uncertainty scores. We then assess if choosing to reject an output improves the overall prediction metric.

Prediction rejection rate (PRR). PRR quantifies how well an uncertainty score U identifies unreliable predictions. It is computed as:

$$\text{PRR} = \frac{\text{AUC}_{\text{uncertainty}} - \text{AUC}_{\text{random}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{random}}},$$

where $\text{AUC}_{\text{uncertainty}}$ is the area under the curve (AUC) obtained by sorting samples by U (highest uncertainty first) and iteratively rejecting the top $N\%$ of samples, computing the average Jaccard score after each rejection threshold. $\text{AUC}_{\text{oracle}}$ is the AUC obtained by sorting samples by the Jaccard score (lowest score first) and rejecting iteratively – this gives us a clear upper-bound to compare against for a given rejection threshold. $\text{AUC}_{\text{random}}$ is the AUC obtained by randomly rejecting $N\%$ of samples at each step. PRR evaluates the effectiveness of U in prioritizing uncertain predictions. A value close to 1 indicates that the uncertainty score performs nearly as well as the oracle, while a PRR near 0 indicates performance no better than random rejection.

Spearman correlation (ρ) of U with correctness. To further assess the quality of the total uncertainty score U, we correlate U with per-sample correctness. We evaluate binary correctness (0/1) for each example–label pair, compute Spearman's ρ between correctness and uncertainty separately for each label, and then report the average ρ across labels. A large negative correlation indicates that higher uncertainty coincides with lower prediction accuracy, confirming that U effectively flags unreliable predictions.

5 RESULTS

We evaluate predictive performance (macro F1, Jaccard J) and uncertainty quantification (PRR, Spearman ρ) across three paradigms: (i) discriminative fine-tuning of encoders/decoders, (ii) generative fine-tuning of decoders, and (iii) 10-shot kNN-based in-context prompting with instruction-tuned or thinking/reasoning models. Labels are grouped by training-set frequency: head (>1%), medium (0.1–1%), tail (0.01–0.1%), and extreme tail (<0.01%). Table 1 summarizes all results.

5.1 MULTI-LABEL PREDICTIVE PERFORMANCE

Fine-tuning. Llama-3.1-8B-Base yields the highest macro F1 overall (0.54), and leads on head (0.74), medium (0.64), and tail (0.53) classes. Smaller Llamas (3.2-3B/1B) and encoders (Ettin-1B/400m/150m) trail slightly, consistent with model capacity, but Ettin-1B-Encoder remains competitive given its size. Llama-3.1-70B-Base, despite being significantly larger, achieves similar results to Llama-3.1-8B-Base in a generative setting. Notably, fine-tuning similarly-sized decoders in a discriminative setting outperforms fine-tuning in a generative setting. We include ablations on LoRA vs. full fine-tuning (Table A.1) and Base vs. Instruct training (Table A.2) in the appendix.

Generative prompting. For prompting setups with 10-shot in-context learning, Qwen3-235B-A22B-Instruct achieves the strongest macro F1 (overall 0.44, extreme tail 0.24), with only GPT-4.1 matching tail performance. Reasoning models excel on the extreme tail (ET): GPT-5 attains the highest ET macro F1 (0.34) and ties the best overall macro F1 (0.54) with Llama-3.1-8B-Base (discriminative). Among open-weight reasoning models, Qwen3-235B-A22B-Thinking is strongest (overall 0.49, tail 0.48, ET 0.33). Gain on rare labels for all prompt-based models comes with a trade-off: head-class performance is consistently below that of the best fine-tuned decoders.

Overall. Discriminative fine-tuning of decoders (Llama-3.1-8B-Base) remains best for head–tail accuracy – this approach consistently outperforms generative training. Prompt-based reasoning models (Qwen3-235B, DeepSeek-R1, GPT-5) dominate the extreme tail and can match overall F1 but lag on head classes. Within prompting, reasoning variants consistently outperform their instruct counterparts (see Qwen3 results).

5.2 UNCERTAINTY QUANTIFICATION

On PRR, across paradigms, generative fine-tuning yields the strongest UQ, followed by discriminative fine-tuning; prompting instruction models ranks third, and reasoning models are a distant last (Table 3). Figure 4 illustrates PRR curves and how PRR summarizes gating quality (shown for Llama-3.1-70B-Base). Overall, Llama-3.2-3B-Base (generative fine-tuning) achieves the highest PRR (0.70) and ρ (-0.46), while thinking models exhibit low PRR ($\leq$ 0.36) and near-zero ρ.

For non-reasoning setups that also deliver strong predictive performance, UQ quality is generally good: Qwen3-235B-A22B-Instruct attains PRR 0.54 and Llama-3.1-70B-Instruct reaches PRR 0.61. Importantly, uncertainty quality does not scale monotonically with model size (e.g., within discriminative fine-tuning, Llama-3.2-1B-Base has the strongest PRR, 0.52, while the 8B variant is lower at 0.47), and the best overall PRR is achieved by a mid-sized generatively fine-tuned model rather than the largest models.

Methodologically, Table 2 shows that the choice of UQ method for a given model has only minor impact on PRR. Further, computing U_{total} yields marginal gains ($\leq$ 0.03 PRR) relative to U_{info} alone, despite incurring roughly 6× compute due to multi-sample generation. Self-verbalized uncertainty performs near chance (PRR = 0), indicating that free-text confidence is not a reliable UQ signal for this task. Finally, closed models that do not expose token-level log-probabilities (e.g., GPT-5) impede UQ and are therefore unsuitable for PRR-based calibration.

Table 1: Predictive performance and uncertainty quantification capabilities of different models and learning paradigms on the truncated test set ($n = 10288$). Macro F1 serves as main predictive metric. Head, medium, tail and extreme tail (ET) classes were assessed separately. For discriminative training, thresholds were chosen for each label on the validation set. J refers to the Jaccard score. PRR is the main metric for assessing uncertainty. For generative models, we report PRR for the best uncertainty quantification method (see U_{info} in 4.3). ρ refers to the Spearman correlation between U and correctness. Up-/downward arrows indicate whether larger (↑) or smaller (↓) values are better. Bold font marks the best model per paradigm. Scores of the best model overall are underlined.

Paradigm/model	**Macro F1 ↑**					**J ↑**	**PRR ↑**	$\rho\downarrow$
	Overall	**Head**	**Medium**	**Tail**	**ET**			
Number of classes →	*1154*	*144*	*481*	*348*	*181*			
Discriminative fine-tuning								
Ettin-150m-Encoder	0.46	0.68	0.56	0.44	0.07	0.55	0.38	-0.30
Ettin-400m-Encoder	0.51	0.72	0.61	0.50	0.12	0.58	0.44	-0.36
Ettin-1B-Encoder	0.53	0.73	0.63	0.51	0.13	0.61	0.46	-0.40
Llama-3.2-1B-Base	0.51	0.71	0.60	0.48	**0.14**	0.58	**0.52**	**-0.42**
Llama-3.2-3B-Base	0.51	0.72	0.62	0.49	0.11	0.59	0.46	-0.41
Llama-3.1-8B-Base	**0.54**	**0.74**	**0.64**	**0.53**	0.12	**0.62**	0.47	-0.40
Generative fine-tuning								
Ettin-400m-Decoder	0.44	0.66	0.54	0.42	0.07	0.54	0.56	-0.44
Ettin-1B-Decoder	0.47	0.67	0.56	0.46	0.10	0.57	0.55	-0.43
Llama-3.2-1B-Base	0.43	0.63	0.52	0.39	0.10	0.45	0.53	-0.44
Llama-3.2-3B-Base	0.48	0.67	0.57	0.46	0.12	0.58	**0.70**	**-0.46**
Llama-3.1-8B-Base	0.50	0.70	0.59	0.48	0.12	0.59	0.43	-0.30
Llama-3.1-70B-Base	**0.53**	**0.73**	**0.62**	**0.51**	**0.16**	**0.61**	0.39	-0.27
Generative prompting – instruct								
Llama-3.1-8B-Instruct	0.08	0.28	0.09	0.03	0.006	0.22	0.22	0.26
Llama-3.1-70B-Instruct	0.30	0.50	0.35	0.25	0.08	0.43	**0.61**	-0.15
Qwen3-4B-Instruct	0.29	0.49	0.35	0.25	0.09	0.41	0.50	-0.27
Qwen3-30B-A3B-Instruct	0.22	0.48	0.27	0.14	0.05	0.43	0.54	0.05
Qwen3-235B-A22B-Instruct	**0.44**	**0.60**	**0.48**	**0.42**	**0.24**	0.49	0.54	**-0.34**
Kimi-K2-Instruct	0.09	0.18	0.11	0.06	0.008	0.07	0.28	0.08
GPT-4.1	0.43	0.59	0.47	**0.42**	0.22	**0.57**	0.43	-0.31
Generative prompting – thinking								
Llama-3.3-Nemotron-49B-v1.5	0.42	0.57	0.46	0.38	0.19	0.46	0.23	-0.03
Qwen3-4B-Thinking	0.38	0.53	0.42	0.36	0.2	0.43	0.18	-0.02
Qwen3-30B-A3B-Thinking	0.45	0.58	0.49	0.44	0.28	0.47	0.09	-0.07
Qwen3-235B-A22B-Thinking	0.49	0.62	0.52	0.48	0.33	0.48	**0.36**	**-0.09**
GLM-4.5-Air	0.42	0.56	0.46	0.39	0.24	0.44	0.24	**-0.09**
DeepSeek-R1-0528	0.48	0.62	0.51	0.47	0.30	0.50	0.26	**-0.09**
GPT-OSS-120B	0.40	0.57	0.45	0.38	0.15	0.45	0.06	-0.0006
GPT-5	**0.54**	**0.68**	**0.58**	**0.53**	**0.34**	**0.57**	NA	NA

6 Discussion

Discussion and conclusion. In this work, we evaluate a range of language models on a newly created unsaturated multi-label text dataset, including different-sized variations of Ettin, LLaMA, Qwen, as well as GLM, Kimi, DeepSeek, and GPT models. Our results show that smaller, task-specific encoder models can achieve state-of-the-art predictive performance while maintaining competitive UQ scores, and additionally offer practitioners flexibility and full control of all model aspects. Large generative fine-tuned models provide slight advantages for underrepresented classes and tend to produce better-calibrated uncertainties. Thinking models in contrast consistently fail to yield a reliable UQ across different model types. Self-verbalized uncertainty in LLMs is highly miscalibrated and should not be used. Prompt-based non-thinking instruct models can produce good but also poor uncertainty scores depending on model type and size, highlighting that fine-tuning is

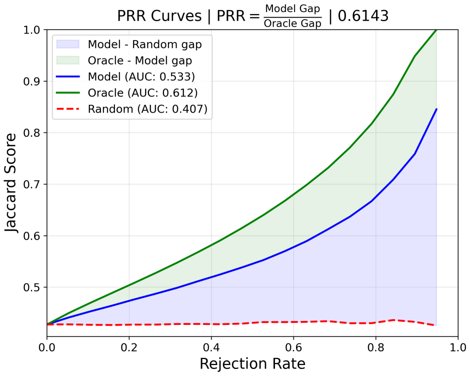

Figure 4: Prediction Rejection Rate (PRR) curves for the Jaccard score for Llama 70B. Three curves represent the model performance under different rejection strategies: oracle (green curve), uncertainty (blue curve), and random (red dashes). Shaded areas indicate the performance gaps between the rejection strategies (blue: 'uncertainty gap' between uncertainty and random; green: 'oracle gap' between oracle and uncertainty). The PRR norm is defined as the ratio of the random gap to the uncertainty gap and indicates the suitability of the uncertainty quantification method (and the model).

Table 2: Comparison of the different UQ methods: information-based (U_{info}), consistency-based (U_{cons}), and information-and-consistency-based ($U_{\text{total}} = U_{\text{info}} * U_{\text{cons}}$) for two of the best performing (non-thinking) generative models. To compute U_{cons}, the models were made to generate 5 outputs at temperature = 1.

Model	**Method**	**PRR**(U_{info})	**PRR**(U_{cons})	**PRR**(U_{total})
Llama-3.1-70B-Instruct	Entropy	**0.60**	-	**0.61**
	Perplexity	0.58	-	0.60
	Max. Token Log Prob.	0.57	-	0.60
	Avg. Logprob	0.59	-	0.60
	Improbability	0.58	-	**0.61**
	Σ EigV. Laplacian	-	0.57	-
	Self-Verbalized	0.00	-	0.36
Qwen3-235B-A22B-Instruct	Entropy	**0.54**	-	**0.57**
	Perplexity	0.53	-	0.56
	Max. Token Log Prob.	0.51	-	0.53
	Avg. Logprob	0.53	-	0.54
	Improbability	0.53	-	0.54
	Σ EigV. Laplacian	-	0.51	-
	Self-Verbalized	0.01	-	0.28

Table 3: Average PRR scores for different U_{info} methods for each generative paradigm from Table 1.

	Entropy	**Perplexity**	**Max. Token Log Prob.**	**Avg. Logprob.**	**Improbability**
Generative fine-tuning	**0.58**	**0.57**	**0.56**	**0.57**	**0.56**
Generative prompting – instruct	0.47	0.44	0.44	0.44	0.45
Generative prompting – thinking	0.20	0.18	0.18	0.18	0.18

generally crucial for obtaining trustworthy uncertainty estimates, for both small discriminative and large generative models. Future work aimed at improving UQ for smaller encoder models could make them the superior choice, combining top-tier predictive performance with trustworthy uncertainty estimates, practical flexibility, and significantly lower computational costs compared to large language models.

Limitations and outlook. Our thorough analysis is carried out on a single newly created dataset. In the future, we want to extend the research conducted in this work to other datasets to further solidify the results. Moreover, in future work, other dimensions of trustworthiness, such as explainability, shall be integrated and studied, to increase the benefits for end-users. Finally, based on our observations, training and developing new reasoning models is a promising direction of research. In fact, we will investigate improving UQ for these models while maintaining high predictive performance across both frequent and infrequent classes.

REFERENCES

Simon Baur, Wojciech Samek, and Jackie Ma. Benchmarking uncertainty and its disentanglement in multi-label chest x-ray classification. *Proceedings of the MICCAI Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, 2025.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Prasoon Varshney, Makesh Narsimhan, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi Mahabadi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Shaona Ghosh, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Chris Alexiuk, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, and Eric Chung. Llama-nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

Lorenzo Bocchi, Camilla Casula, and Alessio Palmero Aprosio. KEVLAR: The complete resource for EuroVoc classification of legal documents. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli (eds.), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pp. 66–73, Pisa, Italy, December 2024. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL https://aclanthology.org/2024.clicit-1.9/.

Guido Boella, Luigi Di Caro, Daniele Rispoli, and Livio Robaldo. A system for classifying multi-label text into EuroVoc. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pp. 239–240, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320801. doi: 10.1145/2514601.2514635. URL https://doi.org/10.1145/2514601.2514635.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6314–6322, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1636. URL https://doi.org/10.18653/v1/P19-1636.

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj, Jingcheng Du, Li Fang, Kai Wang, Shuo Xu, Yuefu Zhang, Parsa Bagherzadeh, Sabine Bergler, Aakash Bhatnagar, Nidhir Bhavsar, Yung-Chun Chang, Sheng-Jie Lin, Wentai Tang, Hongtong Zhang, Ilija Tavchioski, Senja Pollak, Shubo Tian, Jinfeng Zhang, Yulia Otmakhova, Antonio Jimeno Yepes, Hang Dong, Honghan Wu, Richard Dufour, Yanis Labrak, Niladri Chatterjee, Kushagri Tandon, Fréjus A A Laleye, Loïc Rakotoson, Emmanuele Chersoni, Jinghang Gu, Annemarie Friedrich, Subhash Chandra Pujari, Mariia Chizhikova, Naveen Sivadasan, Saipradeep VG, and Zhiyong Lu. Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid track for COVID-19 literature topic annotations. *Database*, 2022:baac069, 08 2022. ISSN 1758-0463. doi: 10.1093/database/baac069. URL https://doi.org/10.1093/database/baac069.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280, 2025. doi: 10.1038/s41467-025-56989-2. URL https://doi.org/10.1038/s41467-025-56989-2.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jinzong Dong, Zhaohui Jiang, Dong Pan, Zhiwen Chen, Qingyi Guan, Hongbin Zhang, Gui Gui, and Weihua Gui. A survey on confidence calibration of deep learning-based classification models under class imbalance data. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (9):15664–15684, 2025. doi: 10.1109/TNNLS.2025.3565159. URL https://dx.doi.org/10.1109/TNNLS.2025.3565159.

Haytame Fallah, Patrice Bellot, Emmanuel Bruno, and Elisabeth Murisasco. Adapting transformers for multi-label text classification. In *CIRCLE (Joint Conference of the Information Retrieval Communities in Europe) 2022*, Jul 2022. URL https://hal.science/hal-03727927v1.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Lukas Galke, Ansgar Scherp, Andor Diera, Fabian Karl, Bao Xin Lin, Bhakti Khera, Tim Meuser, and Tushar Singhal. Are we really making much progress in text classification? a comparative review, 2025. URL https://arxiv.org/abs/2204.03954.

GLM-4.5-Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning

Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intel-*

ligence and Statistics, volume 238 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/harsha-tanneru24a.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proc. EMNLP 2021*, pp. 8153–8161, Online and Punta Cana, Dominican Republic, November 2021. ACL. doi: 10.18653/v1/2021.emnlp-main.643. URL https://doi.org/10.18653/v1/2021.emnlp-main.643.

IMDRF. Adverse Event Terminology — imdrf.org. https://www.imdrf.org/working-groups/adverse-event-terminology, 2025. [Accessed 23-09-2025].

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Kimi-Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.348. URL https://doi.org/10.18653/v1/2024.findings-acl.348.

David D. Lewis. Reuters-21578 text categorization test collection, distribution 1.0. https://www.daviddlewis.com/resources/testcollections/reuters21578/, 1997. [Accessed 23-09-2025].

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=DWkJCSxKU5`.

L Liu, X Yang, J Lei, X Liu, Y Shen, Z Zhang, P Wei, J Gu, Z Chu, Z Qin, et al. A survey on medical large language models: Technology. *Application, Trustworthiness, and Future Directions. arXiv preprint arXiv*, 240603712, 2024.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489. URL `https://doi.org/10.1109/TIT.1982.1056489`.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. URL `https://arxiv.org/abs/1608.03983`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. Large language models do multi-label classification differently, 2025. URL `https://arxiv.org/abs/2505.17510`.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156, 2024. doi: 10.1038/s41598-024-64827-6. URL `https://doi.org/10.1038/s41598-024-64827-6`.

Macedo Maia, Juliano Efson Sales, André Freitas, Siegfried Handschuh, and Markus Endres. A comparative study of deep neural network models on multi-label text classification in finance. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pp. 183–190, 2021. doi: 10.1109/ICSC50631.2021.00039. URL `https://doi.org/10.1109/ICSC50631.2021.00039`.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205. URL `https://doi.org/10.21105/joss.00205`.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020. URL `https://arxiv.org/abs/1802.03426`.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL `https://doi.org/10.18653/v1/N18-1100`.

NeuML. bioclinical-modernbert-base-embeddings. `https://huggingface.co/NeuML/bioclinical-modernbert-base-embeddings`, 2025. URL `https://huggingface.co/NeuML/bioclinical-modernbert-base-embeddings`. Hugging Face model repository.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023. URL `https://arxiv.org/abs/2311.16452`.

Jaya Ojha, Oriana Presacan, Pedro G. Lind, Eric Monteiro, and Anis Yazidi. Navigating uncertainty: A user-perspective survey of trustworthiness of AI in healthcare. *ACM Transactions on Computing for Healthcare*, 6(3):1–32, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

OpenAI. tiktoken. https://github.com/openai/tiktoken, 2022. GitHub repository.

OpenAI. GPT-5 system card. https://openai.com/index/gpt-5-system-card/, 2025a. [Accessed 23-09-2025].

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025b. [Accessed 23-09-2025].

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Tim Schopf, Karim Arabi, and Florian Matthes. Exploring the landscape of natural language processing research. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 1034–1045, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ranlp-1.111/.

Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020. doi: 10.1080/10447318.2020.1741118. URL https://doi.org/10.1080/10447318.2020.1741118.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, June 2025. ISSN 0360-0300. doi: 10.1145/3744238. URL https://doi.org/10.1145/3744238.

Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. JRC Eurovoc indexer JEX - a freely available multi-label categorisation tool. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 798–805, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL https://aclanthology.org/L12-1519/.

Pingjie Tang, Meng Jiang, Bryan (Ning) Xia, Jed W. Pitera, Jeffrey Welser, and Nitesh V. Chawla. Multi-label patent categorization with non-local attention-based graph convolutional network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9024–9031, Apr. 2020. doi: 10.1609/aaai.v34i05.6435. URL https://ojs.aaai.org/index.php/AAAI/article/view/6435.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Uncertainty quantification for LLMs through minimum Bayes risk: Bridging confidence and consistency, 2025. URL https://arxiv.org/abs/2502.04964.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Sida Wang and Christopher Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park (eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90–94, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-2018/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024a. URL https://arxiv.org/abs/2412.13663.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024b. URL https://arxiv.org/abs/2412.13663.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs Seq: An open suite of paired encoders and decoders. *arXiv preprint arXiv:2507.11412*, 2025. URL https://arxiv.org/abs/2507.11412.

Jiageng Wu, Bowen Gu, Ren Zhou, Kevin Xie, Doug Snyder, Yixing Jiang, Valentina Carducci, Richard Wyss, Rishi J Desai, Emily Alsentzer, Leo Anthony Celi, Adam Rodman, Sebastian Schneeweiss, Jonathan H. Chen, Santiago Romero-Brufau, Kueiyu Joshua Lin, and Jie Yang. Bridge: Benchmarking large language models for understanding real-world clinical practice text, 2025. URL https://arxiv.org/abs/2504.19467.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. *arXiv preprint arXiv:2506.21849*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui

Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: Sequence generation model for multi-label classification, 2018. URL https://arxiv.org/abs/1806.04822.

A Appendix

A.1 Dataset topics

Figure A.1 illustrates the topics addressed in the medical device adverse event reports used for testing.

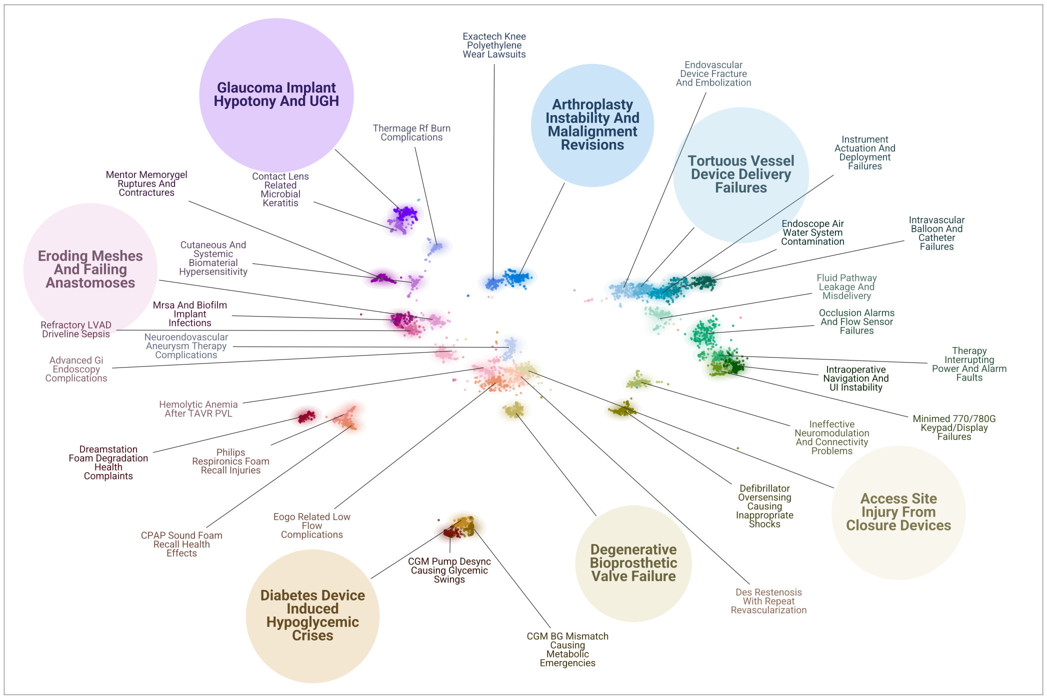

Figure A.1: Illustration of 35 most frequent topics in the (truncated) test set. Topics were modeled with UMAP (McInnes et al., 2020) and K-means clustering (Lloyd, 1982), and named with GPT-5.

A.2 Knowledge cutoff dates

Reports from July 2024 to June 2025 are our test data (see Section 3). This time span ensures that there are no overlaps with the models' pre-training knowledge cutoff dates for Llama 3.1-3 (December 2023), GPT 4.1 (June 2024) and gpt-oss-120b (June 2024). Reports are published *after* each quarter (reports from July-September 2024 in October 2024 or later). Consequently, our test set should also contain only data released after the GPT-5 knowledge cutoff in September 2024. Ettin (Weller et al., 2025) was pre-trained including on the 'DOLMino mix 1124' dataset (OLMo et al., 2025, which also mention data from September 2024), among other data sources. Llama-3.3-Nemotron-49B-v1.5 used post-training data released in July 2025. We have not found publicly available official information on the knowledge cutoff dates of DeepSeek-R1, Qwen3, and Kimi K2 (released in January, April, and July 2025, respectively).

A.3 Hyperparameters

For generative training, we use the AdamW optimizer Loshchilov & Hutter (2019) with cosine learning-rate decay Loshchilov & Hutter (2016), a warmup ratio of 0.1, and a batch size of 64. We tune the learning rate (from 3e-5 to 1e-4) and the maximum gradient norm on the validation set.

For discriminative training, we use AdamW with a cosine scheduler, a fixed batch size of 512, 20 epochs, a warmup ratio of 0.1, and a fixed maximum gradient norm of 0.01; we tune the learning rate in the range [2e-4, 5e-4] on the validation set.

A.4 Additional results

In a generative setting, full fine-tuning is compared to parameter-efficient fine-tuning using LoRA in Table A.1. For Llama models (1B, 3B), full fine-tuning improves macro-F1 by 0.01. For Ettin models (400M, 1B), the gains are greater, at 0.12 and 0.09, respectively. Table A.2 compares base models with instruction-tuned models. Instruction-tuned Llama models (1B, 3B) achieve macro-F1 0.04–0.05 higher than the corresponding base variants.

Table A.1: LoRA vs. full fine-tuning: Performance and UQ capabilities for parameter-efficient (using low-rank adaptation; LoRA) vs. full fine-tuning of generative models (in a generative setting). Macro F1 serves as main predictive metric. Head, medium, tail and extreme tail (ET) classes are assessed separately. J refers to the Jaccard score. PRR is the main metric for assessing uncertainty. We report PRR for the best UQ method (see U_{info} in 4.3). ρ refers to the Spearman correlation coefficient.

Model	**Macro F1 ↑**					**J ↑**	**PRR ↑**	**ρ ↓**
	Overall	**Head**	**Medium**	**Tail**	**ET**			
Number of classes →	*1154*	*144*	*481*	*348*	*181*			
LoRA fine-tuning								
Ettin-400M-Decoder	0.32	0.59	0.41	0.24	0.005	0.48	0.57	-0.45
Ettin-1B-Decoder	0.38	0.63	0.48	0.33	0.014	0.52	0.58	-0.44
Llama-3.2-1B-Instruct	0.45	0.66	0.54	0.44	0.08	0.56	**0.60**	-0.45
Llama-3.2-3B-Instruct	**0.48**	**0.67**	**0.57**	**0.47**	**0.11**	0.58	0.58	**-0.46**
Full fine-tuning								
Ettin-400M-Decoder	0.44	0.66	0.54	0.42	0.07	0.55	0.56	-0.44
Ettin-1B-Decoder	0.47	0.67	0.56	0.46	0.10	0.57	0.56	-0.43
Llama-3.2-1B-Instruct	0.47	0.67	0.56	0.45	0.12	0.57	0.58	-0.43
Llama-3.2-3B-Instruct	**0.49**	**0.68**	**0.57**	**0.48**	**0.14**	**0.58**	**0.59**	**-0.45**

Table A.2: Base vs. Instruct: Performance and UQ capabilities of base models vs. instruction-tuned models, both fine-tuned in a generative setting. Columns named as in Table A.1 above.

Model	**Macro F1 ↑**					**J ↑**	**PRR ↑**	**ρ ↓**
	Overall	**Head**	**Medium**	**Tail**	**ET**			
Number of classes →	*1154*	*144*	*481*	*348*	*181*			
Base model								
Llama-3.2-1B-Base	0.43	0.63	0.52	0.39	0.10	0.45	0.53	-0.44
Llama-3.2-3B-Base	**0.48**	**0.67**	**0.57**	**0.46**	**0.12**	**0.58**	**0.70**	**-0.46**
Instruct model								
Llama-3.2-1B-Instruct	0.47	0.67	0.56	0.45	0.12	0.57	0.58	-0.43
Llama-3.2-3B-Instruct	**0.49**	**0.68**	**0.57**	**0.48**	**0.14**	**0.58**	**0.59**	**-0.45**

A.5 Prompting setup

A.5.1 System prompt

```
SYSTEM_PROMPT = """
You are an AI assistant tasked with classifying a medical device adverse
    event report into one or more categories according to the FDA
    taxonomy. Your goal is to assign all relevant labels to the given
    report.

The report that needs classifying is provided within the <classification-
    text> tag. Along with the report, the label definitions are provided
    within the <labels> tag. To assist you with the task, we also include
     10 "few-shot" examples in the <few-shot-examples> tag. These are
    past reports similar to the one you are classifying - the past
    reports are accompanied by their corresponding labels which were
    tagged by a human expert.

# RULES:

1. The taxonomy of labels is provided within the <labels> tag.
    - Labels are separated by newlines; a definition for the label is
        provided.
    - We are in a 3-level hierarchical multi-label classification setting
         - this means that when a child label (such as A040507) is
        selected, the parent label (A0405) and grandparent label (A04)
        must also be selected. Similarly, if a parent label (A0405) is
        selected, the grandparent label (A04) must also be selected.
    - The converse is not always true - selecting a parent (A0405) or
        grandparent (A04) doesn't necessarily mean selecting all its
        children (A040507).
    - Labels that start with "A" are Medical Device Problems.
    - Labels that start with "E" are Health Effects - Clinical Signs and
        Symptoms or Conditions.
    - The grandparent label (eg. A01) is the most general label, the
        parent label (eg. A0101) is the next most specific, and the child
         label (eg. A010101) is the most specific.
    - The grandparent label always has the letter "A" or "E" followed by
        2 numbers, eg. A01, E01, A02, E02, etc.
    - The parent label always has the letter "A" or "E" followed by 4
        numbers, eg. A0101, E0101, A0201, E0201, etc.
    - The child label always has the letter "A" or "E" followed by 6
        numbers, eg. A010101, E010101, A020101, E020101, etc.

2. There are 10 "few-shot" examples included in the <few-shot-examples>
    tag.
    - Each example includes a report and its corresponding labels.
    - The examples included in the <few-shot-examples> tag were chosen
        using a K-Nearest Neighbours algorithm which picked reports
        similar in content to the text which needs classifying. The
        labels shown for these examples may or may not overlap with the
        labels for the report inside the <classification-text> tag. Use
        them as contextual guidance.

3. Your goal is to classify the text provided within the <classification-
    text> tag.
    - Assign all labels that are relevant.
    - You can choose multiple labels, a single label, or no labels if
        none apply.
    - Always use the exact label names from the label list provided in
        the taxonomy under the <labels> tag.  Do not invent new labels or
         modify existing ones.
    - Return your output as a list of labels, separated by newlines.
    - Do not include any explanations, text, or formatting outside the
        label list.
    - If no label applies, return an empty list.
    - Do not invent new labels.

4. Provide your output as a list of labels, each on a new line. For
    example:

A04
A0405
A040507
E01
E0101

# IMPORTANT
*In your final output, you must not include any extra text, explanations,
     or formatting outside the label list. Only return the list of labels
     separated by newlines.*
"""
```

A.5.2 USER PROMPT

```
USER_PROMPT = """
You are an AI assistant tasked with classifying a medical device adverse
    event report into one or more categories according to the FDA
    taxonomy. Your goal is to assign all relevant labels to the given
    report. The rules were provided in the system prompt. For the sake of
     clarity, we are repeating them here:
- Assign all the labels that are relevant to the report, but only if you
    are sure about it.
- You can choose multiple labels, a single label, or no labels if none
    apply.
- Use exact label names from the provided taxonomy. Do not invent or
    modify labels.
- Include parent and grandparent labels when selecting a child label.
- Do not include any explanations, text, or formatting outside the label
    list.
- If no labels apply, return an empty list.
- Your output should contain only the list of applicable labels, with
    each label on a new line. You must not include any extra text,
    explanations, or formatting outside the label list.
- Provide your output as a list of labels, each on a new line.
Example output:
A04
A0405
A040507
E01
E0101

Here's how to proceed:

1. First,  familiarize yourself with the label definitions:

<labels>
A01: Patient Device Interaction Problem - Problem related to the
    interaction between the patient and the device.
A0101: Patient-Device Incompatibility - Problem associated with the
    interaction between the patient's physiology or anatomy and the
    device that affects the patient and/or the device.
.
.

</labels>

2. Review these few-shot examples of similar reports and their
    corresponding labels:
```

```
<few-shot-examples>
{EXAMPLES}
</few-shot-examples>

3. Now, carefully classify the following report:

<classification-text>
{CLASSIFICATION_TEXT}
</classification-text>"""
```