Where Prompting Hits the Wall: Diagnosing Category Boundary Complexity in LLM-Based Classification

Menglin Liu
The Chinese University of Hong Kong, Shenzhen
menglinliu@cuhk.edu.cn

Xinming Gao

Ge Shi
University of California, Davis

Abstract

Practitioners deploying large language models for classification face a recurring question with no principled answer: when does further prompt engineering stop paying off, and when should one switch to stronger models, supervised training, or human review? We argue that the *escalation behavior* of a staged classification system—how often weaker models disagree and trigger expert or human resolution—is itself a diagnostic signal for this decision. We instantiate this view in PoliPrompt, a framework combining rule-augmented few-shot prompting (RAFS) with a heterogeneous model cascade that produces two operational signals: Expert Call Rate (ECR) and human-in-the-loop (HITL) rate. Across seven text and multimodal datasets, ECR is strongly negatively associated with Macro F1, and 0-shot ECR provides the cleanest rank-ordering of prompting difficulty (Spearman -0.89). This signal supports a three-regime mapping—*near-ceiling*, *prompt-recoverable*, and *prompting-limited*—each implying a different next-step decision. PoliPrompt remains competitive with a GPT-4o single-call baseline on operationalizable tasks and underperforms it mainly on prompting-limited tasks, where high escalation and low F1 indicate that prompting alone is unlikely to suffice. We position PoliPrompt not as a universal replacement for stronger models, but as a diagnostic workflow that tells practitioners when prompting is sufficient, when it is recoverable through structured prompting, and when it has likely hit the wall.

1 Introduction

Large language models have made prompt-based classification widely accessible: given a label set and a few demonstrations, practitioners can obtain usable classifiers without task-specific training (Brown et al., 2020; Liu et al., 2023). This convenience, however, has outpaced our ability to answer a basic deployment question. Practitioners can always engineer one more prompt, retrieve more examples, or add more rules, and recent prompt-programming systems further automate parts of this process (Khattab et al., 2024; Yuksekgonul et al., 2024). Yet current workflows offer little guidance on *when this effort stops paying off*, or when to switch to a stronger model, human review, label redesign, or supervised training.

Standard evaluation does not answer this question. Existing work on calibration and selective prediction studies when individual model predictions may be unreliable (Geifman and El-Yaniv, 2017; Kadavath et al., 2022), but accuracy or F1 after a prompt has been chosen is retrospective: it quantifies how well the current prompt performs, but not *why* it fails, nor whether the failure is recoverable through further prompting. A task that fails because its category boundaries are under-specified is fundamentally different from one that fails because its labels require semantic, pragmatic, or multimodal interpretation that additional prompts cannot easily supply. A single F1 number cannot tell these cases apart.

We argue that the *escalation behavior* of a staged classification system can serve as a diagnostic signal. Building on the intuition behind model cascades and ensemble disagreement (Chen et al., 2023), we treat repeated disagreement among a lightweight model, a heterogeneous verifier, and an expert model as evidence about the task-prompt protocol itself. When these models repeatedly disagree or fail to produce valid labels, this instability reflects how difficult the category boundaries are to resolve under the current task-prompt protocol. We operationalize this behavior through Expert Call Rate (ECR) and human-in-the-loop (HITL) demand. We do not treat ECR as a universal threshold across arbitrary prompts, models, or label sets. Instead, ECR is protocol-relative: under a fixed protocol, its ordering can help practitioners compare a target task against simpler in-protocol anchors or

pilot variants. Across seven datasets, 0-shot ECR is strongly rank-correlated with Macro F1 (Spearman -0.893).

We instantiate this idea in PoliPrompt, a staged prompting framework combining Rule-Augmented Few-Shot prompting (RAFS) with cascaded model arbitration and human correction. Our claim is not that PoliPrompt universally outperforms strong single-call models: it does not, and *where it fails is itself informative*. Across seven text and multimodal datasets, PoliPrompt's escalation signals separate tasks where structured prompting is worthwhile from tasks where prompting appears to hit its ceiling, and this separation aligns with, rather than contradicts, where PoliPrompt underperforms a GPT-4o baseline.

Our contributions are three-fold: **(1)** We formulate *prompting-suitability diagnosis* for classification: the goal is not only to maximize prompt performance, but to decide whether prompt engineering is worth further investment. **(2)** We introduce escalation-based diagnostic signals, especially ECR and HITL demand, and show across seven text and multimodal datasets that they track prompting difficulty; under a fixed protocol, zero-shot ECR provides a strong early-warning ordering of task-prompt instability. **(3)** We instantiate the idea in PoliPrompt and characterize three operational regimes—near-ceiling, prompt-recoverable, and prompting-limited—each implying a different next-step decision for practitioners.

2 Related Work

Prompt-based classification and prompt optimization. Large language models have made prompt-based and few-shot classification practical without task-specific training (Brown et al., 2020; OpenAI, 2024; Liu et al., 2023), and are increasingly applied to domain-specific classification tasks such as financial disclosure analysis (Liu et al., 2026). Prior work studies prompt design, chain-of-thought reasoning (Wei et al., 2022), demonstration selection (Rubin et al., 2022; Su et al., 2023), and retrieval-augmented prompting(Lewis et al., 2020) as ways to improve task performance. These methods mainly ask how to improve the prompt or context. A related PoliPrompt framework has been proposed for political-science text classification, combining enhanced prompt generation, adaptive exemplar selection, and consensus-based arbitration to improve domain-specific classification performance (Liu and Shi, 2024). Our work differs in focus: rather than optimizing a domain-specific classification pipeline, we study prompting-suitability diagnosis and use escalation behavior as a task-prompt-level diagnostic signal across diverse NLP datasets. PoliPrompt instead asks whether continued prompt engineering is worth further investment for a given task-prompt setup.

Uncertainty estimation and selective prediction. Confidence estimation, calibration, uncertainty quantification, and selective prediction study when a model should abstain or defer on individual instances (Geifman and El-Yaniv, 2017; Hendrycks and Gimpel, 2017; Lakshminarayanan et al., 2017; Guo et al., 2017). PoliPrompt operates at a different level. Rather than estimating the reliability of a single prediction, it aggregates disagreement and escalation behavior across a pilot or evaluation set. ECR therefore serves as a task-prompt-level diagnostic signal for deciding whether prompting remains suitable, not merely an instance-level confidence score.

Model cascades and human-in-the-loop systems. Model cascades route inputs through models of increasing cost or capability, often to reduce inference cost while preserving accuracy (Chen et al., 2023). Human-in-the-loop systems similarly defer difficult cases to human annotators. PoliPrompt builds on this routing perspective but uses the routing behavior itself as evidence: expert escalation and HITL demand are not only mechanisms for correcting hard instances, but diagnostic traces of category-boundary instability under the current prompting protocol.

3 PoliPrompt Framework

3.1 Overview

PoliPrompt is designed to produce both task predictions and diagnostic traces for prompt-based classification. Given a dataset, a label set, a task prompt, and a small exemplar pool, the framework first constructs rule-augmented prompting context from the exemplar pool, and then performs staged inference through a heterogeneous model cascade with optional human correction. The output is therefore not only a predicted label for each instance, but also an inference path indicating whether the instance was resolved by lightweight agreement, expert arbitration, or human-in-the-loop correction. These paths are aggregated into diagnostic signals such

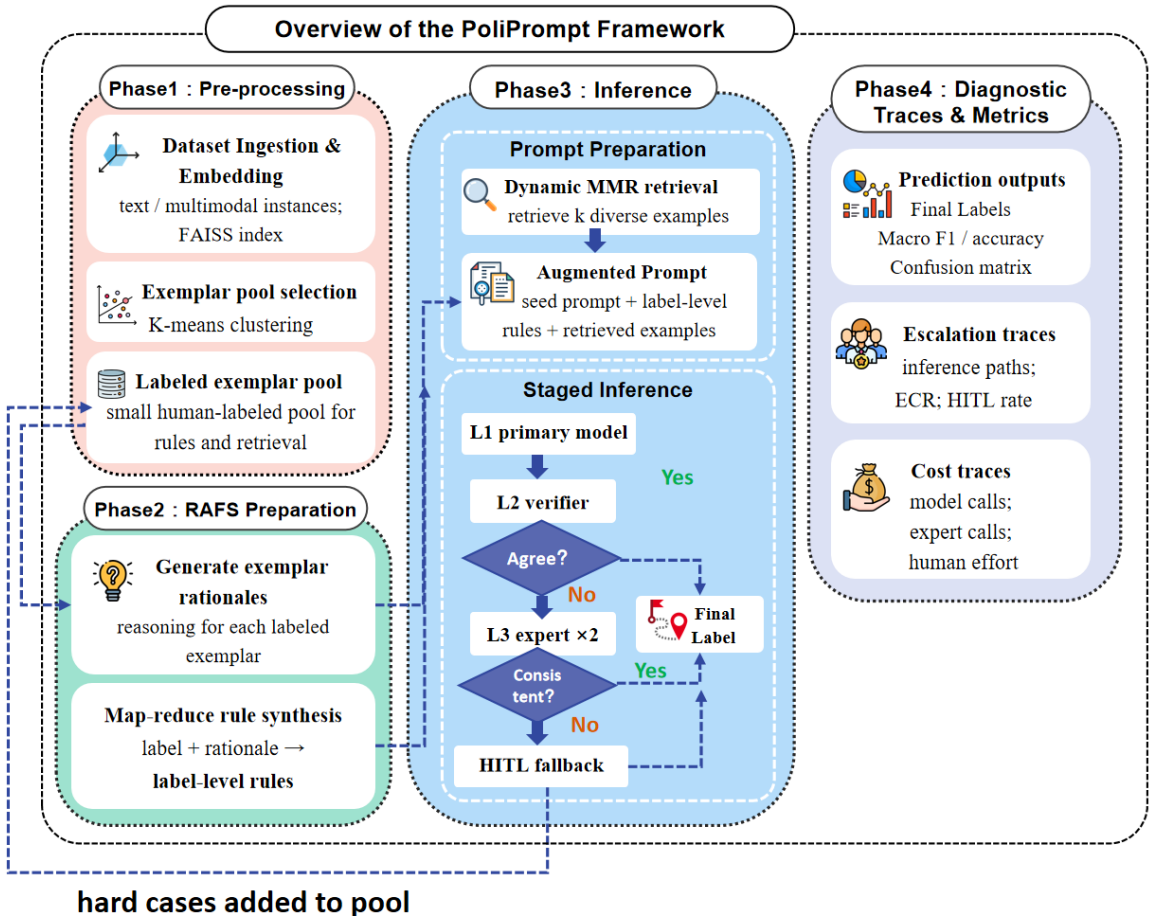

Figure 1: Overview of the PoliPrompt framework. Phase 1 constructs a labeled exemplar pool from the input data. Phase 2 generates exemplar rationales and synthesizes label-level rules through RAFS. Phase 3 assembles query-time augmented prompts through dynamic retrieval and performs staged inference with expert escalation and HITL fallback. The framework produces final labels together with diagnostic traces, including ECR, HITL rate, inference paths, and cost signals.

as Expert Call Rate (ECR) and HITL rate.
[1]

3.2 Rule-Augmented Few-Shot Prompting

Rule-Augmented Few-Shot prompting (RAFS) converts a small labeled exemplar pool into reusable task guidance. PoliPrompt starts from a researcher-provided seed prompt, which specifies the task instruction, label inventory, label definitions, and output format, together with a small human-labeled exemplar pool selected from the dataset.

RAFS produces two artifacts from this exemplar pool. First, for each labeled exemplar, the system generates an *exemplar rationale*: a short explanation of why the instance belongs to its human-provided label. Second, these exemplar rationales are grouped by label and consolidated through a map-reduce style synthesis procedure into *label-level rules*, which summarize category-specific decision criteria, common confusions, and boundary distinctions. This separation is intentional: label-level rules capture cross-instance patterns that may not be visible from any single example, while exemplar rationales preserve local reasoning grounded in specific cases. Both artifacts are stored and reused when constructing augmented prompts at inference time, as described in the next subsection.

3.3 Staged Inference and Escalation

At inference time, PoliPrompt first assembles an augmented prompt for each query. The prompt always includes the researcher-provided seed prompt and the synthesized label-level rules generated by RAFS. In the 0-shot setting, no retrieved instance-level demonstrations are included; in the 5-shot setting, PoliPrompt additionally retrieves $k = 5$ exemplars from the labeled exemplar pool and inserts

[1]Code: https://github.com/rail-ai-lab/poliprompt

their inputs, labels, and exemplar rationales into the prompt. Here, "0-shot" denotes zero retrieved exemplars, not the absence of synthesized rules.

Retrieval is performed with embedding-based search over the labeled exemplar pool, indexed with FAISS, followed by diversity-aware selection using Maximal Marginal Relevance (MMR). MMR balances similarity to the query against similarity to already selected exemplars, with λ controlling the relevance–diversity trade-off. When the shot budget and label space permit, retrieval encourages label coverage; when $k < |\mathcal{Y}|$, the system selects k diverse and relevant exemplars under the fixed budget.

The assembled augmented prompt is then passed through a staged model cascade. A primary model first predicts a label, and a heterogeneous secondary verifier independently evaluates the same instance. The two models are drawn from different model families or providers, so that disagreement is less likely to reflect only sampling noise from a single model. When the primary and secondary predictions match and the label is valid, the instance is resolved without expert intervention.

If the primary and secondary predictions disagree, or if either stage produces an invalid label, the instance is escalated to an expert model. The expert stage is queried twice. If the two expert predictions agree on a valid label, the agreed label is committed as the final prediction. If expert arbitration remains unstable, the instance is passed to human-in-the-loop correction. All HITL-corrected instances are added back to the labeled exemplar pool, so that future retrieval can incorporate these difficult boundary cases. The cascade is provider-agnostic: practitioners can instantiate the primary, verifier, and expert stages with different providers and cost tiers, while keeping the escalation protocol fixed for diagnostic comparison.

Inference as diagnostic trace. The staged design is not only a cost-saving routing mechanism: it produces a per-instance trace of where the task-prompt setup remained stable, as in primary-secondary agreement, and where it broke down, as in expert escalation or HITL correction. The next subsection formalizes this trace into two summary signals.

3.4 Diagnostic Signals

Expert Call Rate and HITL rate. PoliPrompt records an inference path for every processed instance. Let N_{run} denote the number of non-initial-pool instances processed by the system, and let N_{HITL} denote the number of these instances routed to human correction. Since HITL-corrected instances are added back to the exemplar pool, we exclude them from the clean classification evaluation. Let N_{clean} denote the remaining clean evaluated instances after excluding both the initial exemplar pool and HITL-added instances, and let N_{expert} denote the number of clean instances resolved by L3 expert-model arbitration, where two expert calls reach consensus. We report

$$\text{ECR} = \frac{N_{\text{expert}}}{N_{\text{clean}}}, \qquad \text{HITL} = \frac{N_{\text{HITL}}}{N_{\text{run}}}.$$

ECR measures how often the automated cascade requires L3 expert arbitration among clean evaluated cases. HITL rate measures human-correction demand among all non-initial-pool processed instances. We keep the two quantities on different denominators because they measure different operational realities: ECR characterizes instability within the clean evaluation set, while HITL rate characterizes the overall human-correction demand of the deployed system.

These quantities are diagnostic because they summarize where the current task-prompt protocol becomes unstable. The mechanism is structural rather than purely empirical: the primary and secondary models act as heterogeneous voters, and expert escalation is triggered when this agreement test fails. Persistent disagreement therefore provides operational evidence that the category boundaries relevant to the query are difficult to resolve under the current prompt protocol.

We emphasize that ECR and HITL are protocol-relative: their absolute values depend on the choice of primary, secondary, and expert models, the prompt template, the label set, the retrieval setup, and the escalation rule. PoliPrompt therefore uses these signals to characterize instability under a specified protocol rather than to define universal thresholds. We return to practical use without cross-protocol thresholds in Section 7.

4 Experimental Setup

4.1 Datasets

We evaluate PoliPrompt on seven classification datasets spanning both text-only and multimodal settings. The text-only datasets include BBC News for coarse-grained topic classification, CAP Bills for policy topic classification, TweetIrony for pragmatic irony detection, and GoEmotions under the

Ekman emotion mapping. The multimodal datasets include N24News for fine-grained news section classification, CrisisMMD for disaster informativeness classification, and ClimateMemes for climate stance detection. These datasets cover a spectrum of category-boundary conditions, ranging from coarse topic labels to fine-grained ontologies, pragmatic labels, affective labels, and image-text stance reasoning. A safety-sensitive eighth dataset, HatefulMemes, is reported separately in Appendix D.

4.2 Compared Methods

Our primary baseline is a GPT-4o static 5-shot single-call prompting setup. The baseline uses the same final task prompt and label definitions as PoliPrompt, but performs classification in a single model call without dynamic retrieval, RAFS artifacts, staged arbitration, or HITL correction. For multimodal datasets, GPT-4o receives both the textual input and the corresponding image. The static 5-shot baseline selects five demonstrations from the same initial exemplar pool used by PoliPrompt, with exemplar-pool instances excluded from evaluation to avoid leakage. We additionally examine GPT-4o-mini single-call variants on representative datasets as a model-capacity ablation in Section B.

4.3 PoliPrompt Settings

All reported PoliPrompt runs use a fixed staged protocol within each dataset, so that ECR and HITL are interpreted under a stable configuration. Each dataset starts from a researcher-provided seed prompt and an initial human-labeled exemplar pool selected according to dataset size and label cardinality, ranging from 50 to 300 instances. Dataset sizes, sampling details, and clean evaluation denominators are reported in Appendix A.The final prompts are fixed before full-scale runs and shared by PoliPrompt and the single-call baselines.

We evaluate 0-shot and 5-shot settings as defined in Section 3.3; retrieval uses MMR with $\lambda = 0.5$. For text-only datasets, the primary, secondary, and expert stages use GPT-4o-mini, Qwen-Turbo, and GPT-4o, respectively, with text-embedding-3-small for retrieval. For multimodal datasets, the primary and expert stages use GPT-4o-mini and GPT-4o, while the verifier uses Qwen-VL-Plus and retrieval uses qwen3-vl-embedding. Decoding is near-deterministic: temperature 0 for the primary and secondary stages and 0.1 for the two expert calls.

4.4 Evaluation Protocol and Diagnostic Analysis

We use Macro F1 as the primary metric, along with ECR and HITL rate as diagnostic signals. For the main clean evaluation, we exclude both the initial exemplar pool and HITL-added instances, since the latter are incorporated into the exemplar pool after human correction. Classification metrics and ECR are computed on this clean evaluation set. HITL rate is computed separately as the fraction of non-initial-pool processed instances routed to human correction. For the HITL-contribution analysis, we exclude only the initial pool and compare human-corrected predictions with a no-HITL replacement using a single GPT-4o expert pass under the same RAFS context.

We compute Pearson and Spearman correlations between Macro F1 and ECR/HITL across the seven main datasets and visualize the datasets in a diagnostic spectrum. Given the small number of datasets ($N = 7$), we treat these correlations as descriptive evidence of a directional association rather than inferential claims. The resulting signals are protocol-relative indicators of task-prompt instability, not universal task-intrinsic constants.

5 Result

5.1 Overall Performance Across Tasks

Table 1 summarizes the main results across the seven datasets. PoliPrompt's 5-shot setting exceeds GPT-4o on TweetIrony, CAP Bills, and N24News, ties on BBC News and CrisisMMD, and underperforms on ClimateMemes and GoEmotions. Crucially, the two datasets where PoliPrompt clearly underperforms are also the two with the highest 0-shot ECR (29.2% and 22.1%), while the datasets where it gains the most lie in the medium-ECR range. This non-uniformity is central to our diagnostic framing: the purpose of PoliPrompt is not to dominate a strong single-call model on every task, but to make explicit *when* structured prompting is likely to help and *when* it remains insufficient. Section 5.2 shows that escalation behavior tracks prompting difficulty, and Section 5.3 characterizes the boundary conditions under which the cascade is beneficial. We do not compare against supervised systems as primary baselines: our goal is to diagnose whether prompt-based classification is worth further investment, not to replace task-specific training. Reference supervised numbers on five of these datasets are reported separately in

Dataset	0-shot F1	0-ECR	0-HITL	5-shot F1	5-ECR	5-HITL	GPT-4o	GPT-4o-mini
BBC News	0.953	5.3%	0.4%	0.962	3.3%	0.1%	**0.963**	–
TweetIrony	0.812	11.7%	1.0%	**0.826**	15.3%	0.9%	0.808	0.805
CAP Bills	0.706	17.9%	2.1%	**0.724**	13.8%	1.3%	0.669	–
N24News	0.746	17.5%	1.6%	**0.779**	14.6%	2.0%	0.755	0.714
CrisisMMD	0.745	22.8%	2.0%	0.731	18.8%	1.5%	**0.736**	–
ClimateMemes	0.314	29.2%	3.8%	0.468	24.0%	1.1%	**0.589**	–
GoEmotions	0.333	22.1%	1.2%	0.340	21.7%	1.6%	**0.369**	–

Table 1: Main results across seven classification datasets. F1 denotes Macro F1; ECR and HITL are reported as percentages. The GPT-4o column is a static 5-shot single-call baseline using the same final task prompt and label definitions as PoliPrompt. GPT-4o-mini is reported on representative datasets as a model-capacity ablation. Bold indicates the higher score between PoliPrompt's 5-shot Macro F1 and the GPT-4o baseline. "–" indicates the variant was not run. HatefulMemes is excluded from the main analysis and reported separately as a safety-sensitive case (Appendix D).

Appendix E.

5.2 Escalation Signals Track Prompting Difficulty

Figure 2 visualizes the relationship between Macro F1 and ECR across the seven main datasets. Macro F1 is strongly negatively associated with ECR in both inference settings: in 0-shot, the Pearson and Spearman correlations are -0.839 and -0.893; in 5-shot, -0.855 and -0.714. This suggests that ECR captures an operational form of prompting instability: when category boundaries are harder to resolve under a fixed task-prompt protocol, the system more frequently escalates beyond the primary and secondary models.

HITL rate shows the same directional pattern but is noisier than ECR. In 0-shot, Pearson and Spearman correlations between Macro F1 and HITL are -0.642 and -0.786; in 5-shot, they decrease to -0.476 and -0.500. HITL is a sparse final-stage correction signal, whereas ECR captures a broader set of model disagreements and invalid predictions before human intervention. We therefore treat ECR as the primary diagnostic signal and HITL as a residual uncertainty indicator.

The 0-shot setting provides the cleanest rank-ordering signal. While the two settings show comparable Pearson correlation, the Spearman correlation is substantially higher in 0-shot (-0.893) than in 5-shot (-0.714). For early triage, rank consistency is the relevant criterion, since practitioners need to compare task-prompt instabilities relative to one another before investing in retrieval, RAFS, or full-scale inference. This makes 0-shot ECR useful as an early-warning signal: low-ECR tasks such as BBC News (0-shot ECR 5.3%) are already near ceiling; medium-ECR tasks such as CAP Bills (17.9%) and N24News (17.5%) benefit from structured prompting; and high-ECR tasks such as GoEmotions (22.1%) and ClimateMemes (29.2%) expose the limits of prompting-based classification.

Given the small number of datasets ($N = 7$), we interpret these correlations as descriptive evidence of a consistent directional association rather than inferential statistical claims.

5.3 When Does the Cascade Help?

The diagnostic spectrum clarifies when the PoliPrompt cascade is most useful. Its gains are concentrated on tasks where category boundaries are difficult but still operationalizable through rules, retrieved examples, and staged arbitration. CAP Bills (large policy label space), N24News (fine-grained section distinctions), and TweetIrony (literal vs. intended meaning) all fall in this medium-boundary regime, and PoliPrompt outperforms the GPT-4o single-call baseline on each. On BBC News, all methods are near ceiling. On CrisisMMD, 5-shot retrieval slightly degrades performance (5-shot 0.731 vs. 0-shot 0.745), consistent with the instance-level variance from retrieved exemplars discussed in Section 5.2.

The clearest negative cases are ClimateMemes and GoEmotions, both high-ECR datasets. ClimateMemes often requires identifying the target of satire, resolving multimodal stance, and interpreting cultural context; GoEmotions under the Ekman mapping compresses diverse Reddit affective expressions into broad categories, creating fuzzy label boundaries. In such cases, persistent high escalation combined with low F1 indicates that additional prompting structure may not be sufficient—stronger models, label redesign, task-specific train-

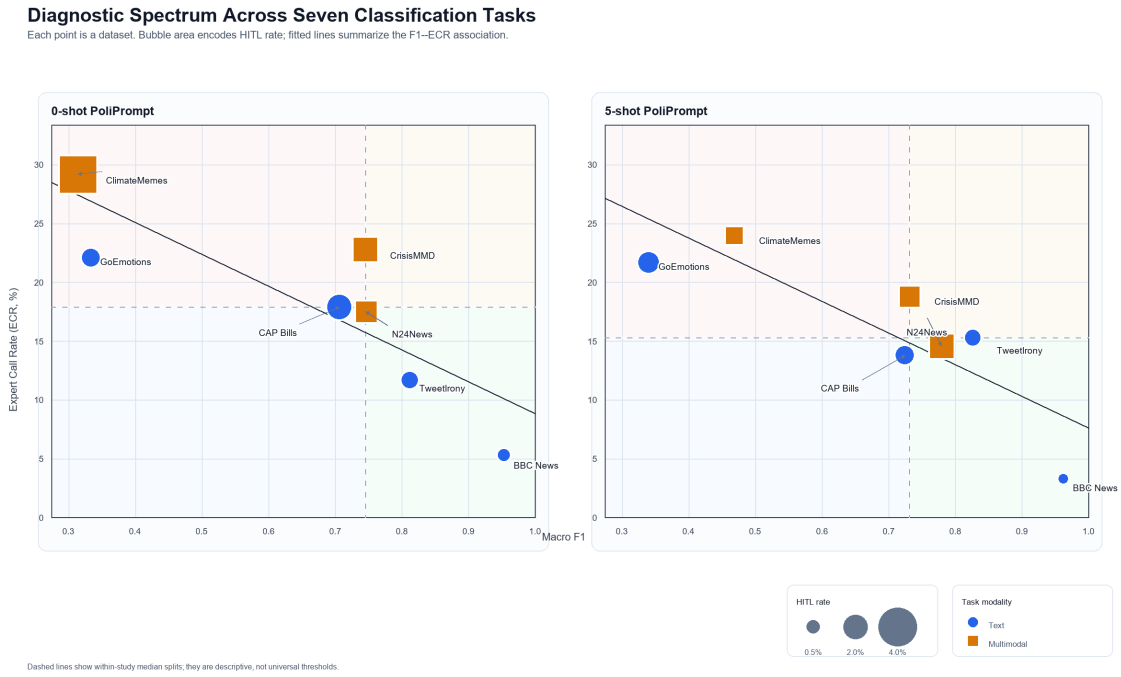

Figure 2: Diagnostic spectrum across seven classification tasks. Each point is a dataset; bubble area encodes HITL rate and marker style denotes modality. Across both 0-shot and 5-shot settings, higher Macro F1 tends to coincide with lower Expert Call Rate (ECR). Dashed lines indicate within-study median splits and are descriptive rather than universal thresholds.

ing, or more substantial human supervision may be more appropriate.

Additional ablations on model capacity, RAFS components, HITL contribution, and prompt refinement are reported in Appendix B; these results are consistent with the diagnostic interpretation above.

5.4 Failure Analysis

The high-ECR, low-F1 cases clarify where prompting fails. ClimateMemes and GoEmotions require semantic, pragmatic, or affective judgments that are difficult to operationalize through additional rules and demonstrations alone, suggesting that stronger models, label redesign, or task-specific training may be more appropriate. A separate operational failure mode arises on safety-sensitive content, where provider-side moderation can intercept inputs before classification; we discuss this case in Appendix D.

6 Limitations

We acknowledge three limitations of this study:

- **Protocol-relative signals.** ECR and HITL are operational signals whose absolute values depend on the prompt template, model cascade, label set, retrieval setup, and escalation rule. We do not propose universal cross-protocol thresholds; practitioners should interpret ECR through in-protocol self-calibration (Section 7). Different seed prompts or label definitions may shift ECR, so the signal diagnoses a task-prompt setup rather than a task in isolation.

- **Dataset coverage.** Our empirical analysis covers seven main datasets. This is sufficient to reveal a consistent directional association between escalation behavior and prompting difficulty, but we treat the correlations as descriptive rather than inferential. Larger and more diverse task collections are needed for population-level claims.

- **Comparison scope.** PoliPrompt is not designed to replace supervised fine-tuning when large labeled training sets are available. Its role is earlier-stage triage: diagnosing whether prompt-based classification is sufficient, recoverable through structured prompting, or likely to require stronger supervision, label redesign, or human review.

7 Discussion

What ECR diagnoses, and what to do next. ECR is most useful when interpreted together with Macro F1. Across our seven datasets, three operational regimes emerge:

- **Low ECR with high F1** (e.g., BBC News, 0-shot ECR 5.3% with F1 0.953): a *near-ceiling regime*. The current prompt-based setup is stable, and additional prompt engineering is unlikely to yield large gains.
- **Medium ECR with medium-to-high F1** (e.g., CAP Bills, N24News, and TweetIrony, with 0-shot ECRs around 11%–18%): a *prompt-recoverable regime*. Category boundaries are difficult but still operationalizable; structured prompting through label-level rules and retrieved exemplars can improve performance. Demo-scale prompt refinement on TweetIrony and N24News further supports that this regime is responsive to clearer boundary specification.
- **High ECR with low F1** (e.g., GoEmotions and ClimateMemes, with 0-shot ECRs 22.1% and 29.2%): a *prompting-limited regime*. Persistent disagreement suggests that the bottleneck is unlikely to be prompt wording alone, but may involve deeper semantic, pragmatic, affective, or multimodal interpretation. Stronger models, supervised fine-tuning, label redesign, or more substantial human supervision may be more appropriate next steps.

This mapping turns ECR into actionable guidance: it indicates not only that a task is unstable under prompting, but what kind of intervention is more plausible next.

Protocol-relative self-calibration. ECR should not be interpreted as a universal threshold. A 15% or 20% ECR under one model cascade is not automatically comparable to the same value under another cascade, prompt template, or label set. The key evidence is rank-based rather than threshold-based: under a fixed protocol, ECR provides a strong task-level ordering signal (Spearman -0.893 in 0-shot), which is what early triage requires. For a single target task, practitioners should fix the prompt template, label definitions, model cascade, retrieval configuration, and escalation rule, then compare the target task against an in-protocol easy anchor, such as a coarse-label version, a clearly separable subset, or a small set of high-confidence examples. The target task's regime should be identified through in-protocol comparison, not by applying our numerical cutoffs directly.

Relation to supervised systems. PoliPrompt is a triage tool, not a replacement for supervised learning. On easier or more operationalizable tasks, prompt-based workflows can avoid task-specific training. On high-ECR, low-F1 tasks, stronger models, supervised fine-tuning, label redesign, or additional human review may be necessary. Reference supervised results in Appendix E provide context but are not primary baselines, since they differ in training data, splits, preprocessing, and label mappings.

Broader implications. This framing aligns with a broader need to rethink evaluation in the era of widely available general-purpose language models. The question is not only which prompt obtains the best benchmark score, but whether the data, labels, prompts, and deployment protocol can support reliable prompt-based classification in the first place. Diagnostic signals such as ECR shift evaluation from static performance leaderboards toward operational decisions under uncertainty.

8 Conclusion

This paper introduced prompting-suitability diagnosis for classification: the practical question is not only how well a prompt performs, but whether continued prompt engineering remains worthwhile. We instantiate this idea in PoliPrompt, a staged prompting framework that combines RAFS with heterogeneous model arbitration and produces both predictions and escalation traces. Across seven text and multimodal datasets, ECR is strongly negatively associated with Macro F1, and 0-shot ECR provides the cleanest rank-ordering of task-prompt instability. The results support a three-regime view: near-ceiling tasks where additional prompting offers limited returns, prompt-recoverable tasks where structured prompting helps, and prompting-limited tasks where stronger supervision, label redesign, or human review may be needed. PoliPrompt is therefore best viewed as a diagnostic workflow for deciding when prompting is sufficient, still recoverable, or likely to have hit the wall.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113. Association for Computational Linguistics.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Menglin Liu and Ge Shi. 2024. PoliPrompt: A high-performance cost-effective LLM-based text classification framework for political science. *arXiv preprint arXiv:2409.01466*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yue Liu, Zhiyuan Cheng, and Longying Lai. 2026. Improving the completeness and comparability of segment disclosures: A large language model approach. *Available at SSRN 6720239*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Association for Computational Linguistics.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. In *Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management*. ISCRAM.

OpenAI. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Ji Xin, Rui Zhang, Mari Ostendorf, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775. European Language Resources Association.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. TextGrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2023. PromptBench: A unified library for evaluation of large language models. *arXiv preprint arXiv:2312.07910*.

A Dataset and Evaluation Statistics

Table 2 reports the number of instances used in each experiment and the clean evaluation denominators after excluding the initial exemplar pool and HITL-added instances. BBC News uses the full dataset; TweetIrony, CrisisMMD, and ClimateMemes use official test or evaluation subsets; CAP Bills, GoEmotions, and N24News use stratified evaluation subsets to ensure label coverage and control evaluation cost.

B Ablations and Diagnostic Utility

We isolate three components of the framework: model capacity, rule augmentation (RAFS), and human-in-the-loop correction (HITL). We also analyze whether demo-stage prompt refinement improves tasks that our diagnostic framing identifies as prompt-suitable or boundary-recoverable. The model-capacity ablation is evaluated on full-scale runs with the final prompts, while the RAFS and prompt-refinement analyses use matched demo subsets. HITL is analyzed on full N24News because demo-scale runs contain too few HITL-triggered instances for a meaningful comparison.

Dataset	Total	$\lvert\mathcal{Y}\rvert$	Pool	0-shot	5-shot
BBC News	2225	5	100	2116	2123
CAP Bills	3000	20	300	2644	2664
TweetIrony	784	2	100	677	678
GoEmotions	1639	7	100	1520	1515
N24News	2354	24	300	2022	2013
CrisisMMD	2237	2	100	2095	2104
ClimateMemes	235	3	50	178	183

Table 2: Dataset sizes, label counts, initial exemplar-pool sizes, and clean evaluation denominators. "Total" denotes the number of instances used before excluding exemplar-pool and HITL-added instances. "Pool" denotes the initial human-labeled exemplar pool. "0-shot" and "5-shot" denote clean evaluation counts under each setting.

Dataset	Earlier	Final	w/o RAFS
TweetIrony	0.776	**0.870**	0.833
N24News	0.763	**0.786**	0.764

Table 3: Demo-scale prompt-refinement and RAFS ablation results. Scores are Macro F1 on matched demo subsets. "Earlier" denotes a development-stage prompt; "Final" denotes the boundary-explicit prompt used in the main experiments; "w/o RAFS" removes rule-augmented few-shot rationales under the final protocol.

Model capacity. Replacing the full cascade with a single GPT-4o-mini call yields a consistent ordering on N24News: Macro F1 increases from 0.714 with GPT-4o-mini to 0.755 with GPT-4o and 0.779 with PoliPrompt. This suggests that both stronger base models and the cascade contribute on fine-grained multimodal classification. On TweetIrony, the differences are smaller (0.805, 0.808, and 0.826, respectively), indicating that once the decision boundary is clearly specified, even a smaller single-call model can perform competitively. This is consistent with our diagnostic framing: structured prompting is most useful when boundaries are difficult but still operationalizable.

Prompt refinement and RAFS. Table 3 shows that prompt refinement and RAFS both contribute to boundary specification on matched demo subsets. Replacing earlier task prompts with the final boundary-explicit prompts improves TweetIrony from 0.776 to 0.870 and N24News from 0.763 to

0.786. These development observations support our diagnostic interpretation: when a task falls in the prompt-suitable or boundary-recoverable region, clearer prompt-level boundary specification can yield substantial gains. Removing RAFS under the final-prompt protocol reduces TweetIrony from 0.870 to 0.833 and N24News from 0.786 to 0.764, suggesting that rule-augmented rationales provide an additional, moderate benefit beyond the prompt template itself.

Human-in-the-loop. We assess the HITL contribution on N24News, which has a large label space and sufficient HITL-triggered cases for analysis. Under the HITL-analysis protocol described in Section 4.4, replacing every human-corrected decision with a single GPT-4o expert pass under the same RAFS context reduces Macro F1 from 0.7855 to 0.7716 (-0.0139). The drop is small in absolute terms because HITL is triggered on only about 2% of instances, but it is non-negligible relative to the gains from other components. HITL therefore functions as a sparse, targeted correction on the hardest residual cases rather than a brute-force source of accuracy.

C Additional Diagnostic Correlations

Figure 3 reports Pearson and Spearman correlations between Macro F1 and the diagnostic signals (ECR, HITL) across the seven main datasets, in both 0-shot and 5-shot settings. The pattern is consistent with the analysis in Section 5.2: ECR shows a stronger negative association with Macro F1 than HITL, and 0-shot ECR provides the cleanest rank-ordering signal.

D Safety-Sensitive Case Study: HatefulMemes

This appendix reports the safety-filtering observations on HatefulMemes, which we exclude from the main seven-dataset analysis (Section 5.4). We report HatefulMemes separately because its escalation behavior is confounded by provider-side safety filtering. Unlike the seven datasets in the main analysis, where escalation primarily reflects model disagreement, invalid labels, or unresolved category boundaries, HatefulMemes contains safety-sensitive content that can trigger moderation filters during inference. These refusals produce forced failures at the model-call level and therefore cannot be interpreted as ordinary evidence of category-boundary ambiguity.

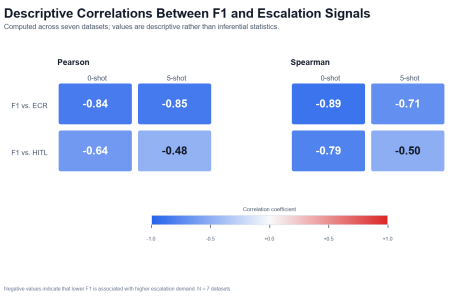

Figure 3: Descriptive correlations between Macro F1 and escalation signals across the seven main datasets. ECR shows a stronger negative association with F1 than HITL in both 0-shot and 5-shot settings. The 0-shot setting yields the clearest rank-ordering signal, as reflected in the stronger Spearman correlation between F1 and ECR. All values are descriptive rather than inferential statistics.

In our HatefulMemes setup, the L2 verifier (Qwen-VL-Plus) sometimes returned a content-inspection error instead of a valid classification label. In the 5-shot setting, 67 out of 500 instances (13.4%) were intercepted during L2 inference. In the 0-shot setting, 15 out of 500 instances (3.0%) were intercepted. Table 4 summarizes these safety-filtering failures.

Setting	Total	Filtered	Rate
0-shot	500	15	3.0%
5-shot	500	67	13.4%

Table 4: Provider-side safety-filtering failures observed on HatefulMemes. These failures occurred during L2 inference and are treated separately from ordinary expert escalation.

This case highlights an important boundary of our diagnostic framework. ECR and HITL are intended to measure operational instability under a fixed task-prompt protocol, but safety-filtering failures introduce a different source of instability: the model provider may refuse to process the input before the classification decision is made. In such cases, escalation traces conflate semantic uncertainty with deployment constraints. We therefore exclude HatefulMemes from the main seven-dataset correlation and spectrum analyses. On safety-sensitive tasks, ECR and HITL should be interpreted alongside refusal rates rather than as standalone diagnostic signals.

E Reference Published Results

For context, we report representative published or supervised reference results from prior work on a subset of our datasets. These numbers are not used as primary baselines, because they differ from our prompt-only setting in training data availability, evaluation splits, label mappings, model inputs, and metric definitions. We therefore treat them as contextual reference points rather than direct competitors. We omit datasets for which we could not identify a sufficiently comparable published reference under matched conditions. In particular, BBC News, CAP Bills, and ClimateMemes are omitted because available references either use different task formulations or do not provide a directly comparable supervised fine-tuned reference under matched conditions.

Dataset	Reference	Metric	PoliPrompt
TweetIrony	0.821	F1 (irony)	0.826
N24News	0.825	F1	0.779
GoEmotions	0.640	Avg. F1	0.340
CrisisMMD	0.842	Binary F1	0.731

Table 5: Representative published reference results. Values are contextual rather than directly comparable.

TweetIrony uses the BERTweet result(Nguyen et al., 2020) on the TweetEval irony task (Nguyen et al., 2020; Barbieri et al., 2020). N24News uses the Image+Abstract multimodal result reported by Wang et al. (2022). GoEmotions uses the BERT Ekman reference reported by Demszky et al. (2020). CrisisMMD uses the informativeness reference reported by Ofli et al. (2020). These values may differ in training regime, split, label mapping, modality, and metric definition, and are therefore used only as contextual references.

The reference results are consistent with our diagnostic framing, but should not be interpreted as a direct leaderboard comparison. On TweetIrony, PoliPrompt is close to the BERTweet reference result, consistent with our finding that relatively clear pragmatic boundaries can be operationalized through prompting, rules, and retrieved examples. On N24News, PoliPrompt remains below the published multimodal reference but within a moderate range, matching our interpretation that the task is difficult but still boundary-recoverable. On GoEmotions and CrisisMMD, published supervised or multimodal systems retain a larger advantage, supporting our claim that high-escalation or semantically demanding tasks may require supervised training, richer supervision, or label redesign rather than additional prompt engineering alone.

Method family	Goal	Unit	Task diag.
Prompt toolkits	Build classifiers	Task/model	No
Prompt benchmarks	Evaluate prompts/LLMs	Task/model	Partial
Prompt optimization	Improve prompts	Prompt/task	No
Selective prediction	Abstain selectively	Instance	No
PoliPrompt	Diagnose suitability	Task-protocol	Yes

Table 6: Positioning of PoliPrompt relative to existing prompt and uncertainty methods.

F Positioning Relative to Existing Tools

PoliPrompt is complementary to prompt-learning, prompt-benchmarking, prompt-optimization, and uncertainty-estimation methods. Existing tools primarily help practitioners build prompts, evaluate prompts or models, improve prompts, or abstain on individual instances. PoliPrompt instead diagnoses whether continued prompt engineering is likely to be worthwhile for a task-prompt protocol.

Prompt-learning toolkits such as OpenPrompt (Ding et al., 2022) help practitioners construct prompt-based classifiers. Prompt benchmarks such as PromptBench (Zhu et al., 2023) and BIG-bench (Srivastava et al., 2023) evaluate prompt or model performance. Prompt-optimization methods such as DSPy/MIPRO (Khattab et al., 2024; Opsahl-Ong et al., 2024) and TextGrad (Yuksekgonul et al., 2024) are useful after a practitioner decides to keep improving a prompt, but they do not directly answer whether further prompt engineering is worthwhile. Selective-prediction methods (Geifman and El-Yaniv, 2017) decide which individual instances to abstain on, whereas PoliPrompt aggregates escalation behavior across instances to expose task-level instability before further prompt engineering, stronger models, or supervised training.

G Prompt Templates

We provide representative prompt templates and RAFS artifacts rather than full raw prompts for every dataset. The examples below illustrate the common prompt structure used across tasks and the main differences between zero-shot and RAFS-augmented five-shot inference.

Dataset	$\lvert\mathcal{Y}\rvert$	# Rules
BBC News	5	5
ClimateMemes	3	10
TweetIrony	2	12
CrisisMMD	2	14
GoEmotions (Ekman-7)	7	37
CAP Bills	20	132
N24News	24	182

Table 7: Enhanced rule-book size across the seven datasets in our main experiments. Rule counts are measured after RAFS rule synthesis and reflect both label-space cardinality and the amount of boundary clarification encoded by the generated rule book.

G.1 Generic Zero-shot Template

The zero-shot setting uses the task prompt, label definitions, and synthesized rules, but does not include retrieved demonstrations. A representative template is:

```
You are a classification system. Your task is
    to classify the input into one of the
    predefined labels.

# LABEL DEFINITIONS
[Dataset-specific label definitions]

# SYNTHESIZED RULES
[Label-level rules generated by RAFS]

# INSTRUCTION
Analyze the input carefully and choose the
    single best label.

# INPUT
[Current instance]

# FORMAT REQUIREMENT
Return your response as a JSON object with
    exactly two fields:
{
  "label": "...",
  "reason": "..."
}
```

G.2 Generic RAFS Five-shot Template

The five-shot setting augments the same task prompt with retrieved demonstrations and exemplar rationales. The retrieved examples are selected from the labeled exemplar pool using MMR to balance relevance and diversity.

```
You are a classification system. Your task is
    to classify the input into one of the
    predefined labels.

# LABEL DEFINITIONS
[Dataset-specific label definitions]

# SYNTHESIZED RULES
[Label-level rules generated by RAFS]

# RETRIEVED EXAMPLES
Example 1:
Input: ...
Label: ...
Rationale: ...

Example 2:
Input: ...
Label: ...
Rationale: ...

...

# CURRENT TASK
Input: [Current instance]

# FORMAT REQUIREMENT
Return your response as a JSON object with
    exactly two fields:
{
  "label": "...",
  "reason": "..."
}
```

G.3 TWEETIRONY Excerpt

TWEETIRONY is a binary pragmatic classification task. The final prompt makes the boundary between literal and intended meaning explicit:

```
# LABEL DEFINITIONS
"0" (non-ironic): The tweet should be
    interpreted literally. The author means
    exactly what they say.

"1" (ironic): The tweet expresses the opposite
    of its literal meaning, often using praise
    to criticize, complain, or convey
    frustration.

# KEY SIGNALS
- Polarity contrast between literal wording and
     intended meaning.
- Exaggerated praise for something negative or
    unpleasant.
- Complaint or criticism disguised as
    positivity.
- Rhetorical wording that should not be
    interpreted literally.

# NEGATIVE CRITERION
Do not label a tweet as ironic merely because
    it is emotional, humorous, exaggerated, or
    informal.
```

This excerpt illustrates how prompt refinement makes pragmatic decision boundaries more operational. In the matched demo subset used for development, the refined prompt improved TWEETIRONY from 0.776 to 0.870 Macro F1. These refinement results are measured on matched demo subsets and are not the main test-set results reported in Table 1.

G.4 N24NEWS Excerpt

N24NEWS has a large label space, so the prompt emphasizes pairwise distinctions between semantically adjacent sections. The following excerpt shows part of the final boundary-explicit prompt:

```
# KEY DISTINCTIONS

- Fashion & Style:
  Choose this label for fashion industry,
      clothing,
  designers, runway shows, luxury brands, beauty
      ,
  celebrity outfits, or fashion events.

- Style:
  Choose this label for social behavior,
      etiquette,
  relationships, identity, lifestyle trends,
      personal
  conduct, social norms, and how people present
  themselves. Do not use this label merely
      because
  the article is about fashion products.

- Health:
  Choose this label for medical issues,
      diseases,
  treatments, public health, clinical research,
       drugs,
  addiction, healthcare, or health policy.

- Well:
  Choose this label for personal wellness,
      fitness,
  nutrition, mental well-being, mindfulness,
      habits,
  self-care, family health, and practical
      lifestyle
  advice for living healthier.
```

This excerpt illustrates how the final prompt adds boundary-specific distinctions rather than only listing labels. In the matched demo subset used for development, the refined prompt improved N24NEWS from 0.763 to 0.786 Macro F1. These refinement results are measured on matched demo subsets and are not the main test-set results reported in Table 1.

G.5 Prompt Refinement Summary

The matched demo-subset prompt-refinement results used during development are summarized below. These are not the main test-set results in Table 1.

Dataset	Earlier	Final	Δ
TweetIrony	0.776	**0.870**	+0.094
N24News	0.763	**0.786**	+0.023

TweetIrony refinement broadened the non-ironic definition, added explicit negative criteria, and enumerated irony signals. N24News refinement expanded pairwise label distinctions, added primary-topic tie-breakers, and introduced a negative criterion for *Opinion*.