The Selection Effect: Why Compact LLM Verifiers for Emotion Classification Fail Behind Encoder Confidence Gates

Dmitry Tikhomirov[1], **Andrey Savchenko**[2,3]
[1]Central University, [2]Sber AI Lab, Moscow, Russia,
[3]HSE University, Laboratory of Algorithms and Technologies for Network Analysis

Abstract

Is a compact LLM verifier cost-effective for closed-label emotion classification when a supervised encoder is already available? We formulate inference-time escalation as cost-aware selective routing: an encoder predicts first, and only uncertain examples (via confidence, disagreement, or novelty gates) may be forwarded to a verifier. We run a Pareto search over thresholds and seven verifier slots, including zero-shot Qwen3.5-4B, LoRA-tuned Qwen, voting, self-consistency, MultiAgentESC-style routing, heterogeneous voting, and a second encoder. Across three-seed RoBERTa experiments, DeBERTa robustness checks, and an additional Russian XLM-R diagnostic, no compact-LLM verifier improves the high-quality low-cost frontier. The mechanism is a selection effect: encoder gating forwards a conditional slice on which compact LLMs perform no better than the encoder itself, or substantially worse, erasing any standalone LLM advantage. Encoder-only gating, split-conformal acceptance, and a small RoBERTa ensemble are strictly preferable alternatives.

1 Introduction

LLM-based multi-agent systems increasingly use compact language models as verifiers to correct supervised models, trading inference cost for quality (Chen et al., 2024a; Yuan et al., 2025; Zhang et al., 2025a), yet agentic workflows can incur substantial overhead relative to strong single-agent baselines (Xu et al., 2026). This raises a basic question: *when a supervised encoder is already available, does a compact LLM actually improve the cost-quality frontier?* We study it in closed-label emotion classification (Demochkina and Savchenko, 2021; Lapokin and Savchenko, 2026; Mozikov et al., 2024a; Zhang et al., 2025b), where the label space is small and the encoder can be trained directly on task data. By compact, we mean inference-time verifiers in the 3-4B parameter range, motivated by their intended role as low-overhead correction models. The answer is not obvious: compact LLMs can exploit label descriptions and general linguistic knowledge without supervised training, while a supervised encoder has a strong inductive advantage on a fixed label set. We hypothesize that the *uncertainty gate itself* changes this comparison. An encoder forwards precisely the examples on which it is uncertain, but encoder uncertainty need not coincide with LLM-correctability. Thus, an LLM that is useful as a full-coverage classifier may provide little or no benefit on the conditional slice selected for verification.

We formalize verification as cost-aware selective routing: the encoder predicts first, and only uncertain examples may be abstained from or forwarded to an optional verifier, with NSGA-II searching over confidence functions, thresholds, and seven verifier configurations. So that the answer cannot hinge on one trigger, every verifier is re-evaluated under six alternative escalation gates and an error oracle. Our experiments reveal a consistent selection effect. Across three encoder seeds, two encoder families, two LLM families, and multiple verifier designs, no compact LLM improves the high-quality, low-cost frontier: on the forwarded slice, the best compact verifier ties the encoder and the rest are substantially worse. That slice is enriched for ambiguous, minority-class, and novel examples, where encoder and verifier errors track human annotation agreement. Scaling the verifier from 4B to 397B parameters does not close the gap, indicating that the limitation is conditional task alignment rather than simply model capacity.

This paper makes three contributions[1]. First, we formulate compact-LLM verification as a cost-aware selective-routing problem, explicitly placing

[1]Code, cached verifier predictions and frozen splits are available at https://github.com/tikhomirovd/selective-emotion-classification.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 24095–24114
October 24-29, 2026 ©2026 Association for Computational Linguistics

abstention and LLM escalation in competition. Second, we establish a robust negative result across encoder seeds, model families, verifier designs, and routing gates: compact LLM verifiers do not improve the high-quality, low-cost frontier when an in-domain supervised encoder is available. Third, we identify the mechanism behind this result: uncertainty gating selects a conditional slice characterized by ambiguity and novelty on which same-taxonomy LLM verifiers provide little information.

The rest of the paper is organized as follows. The proposed framework is presented in Section 2. Main experimental results are discussed in Section 3. Our mathematical notation is summarized in Appendix A. Related works are discussed in Appendix B. Appendix C contains details of experimental settings, while additional experimental results are presented in Appendix D.

2 Methodology

We formulate inference-time verification as a *cost-aware selective-routing problem*. Given a fine-tuned encoder, each input is either accepted using the encoder prediction, abstained on, or forwarded to a verifier. We search over confidence rules, thresholds, and verifier choices to test whether compact LLM verification improves the Pareto frontier.

Routing procedure. Let f_θ produce logits $\ell \in \mathbb{R}^K$ over a closed Ekman-style ($K = 7$) emotional label set (Ekman, 1972). We obtain temperature-scaled probabilities $p = \mathrm{softmax}(\ell/T)$, where temperature T is fitted on the validation set. This calibration is used to separate uncertainty estimation from the arbitrary scale of the raw softmax outputs (Guo et al., 2017). Calibration diagnostics before and after scaling are detailed in Appendix C.6.

Next, we define the encoder prediction as

$$\hat{y} = \arg\max_y p_y, \tag{1}$$

and compute a confidence score $\gamma = \phi(p) \in [0, 1]$ using one of three standard functions: the probability margin ($p^{(1)} - p^{(2)}$), the maximum probability ($p^{(1)}$), or the entropy-based confidence ($1 - H(p)/\log K$). We include all three in the search rather than selecting one a priori; the margin score is selected most often and is analyzed further in Appendix D.11.

For a confidence-specific threshold $\theta_c \in [0.30, 0.99]$, examples from the accepted set $A_c = \{x : \gamma(x) > \theta_c\}$ are accepted and assigned the encoder prediction $\hat{y}$. This is the standard selective-prediction routing mechanism (Chow, 1970; El-Yaniv and Wiener, 2010).

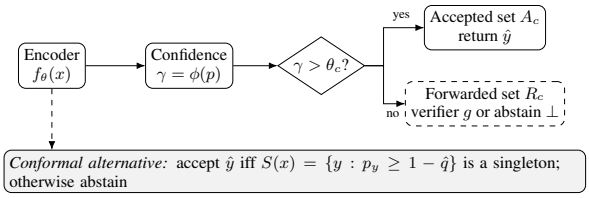


Figure 1: Selective routing with an optional compact LLM verifier. The encoder first produces a prediction and confidence score. Examples above the threshold form the accepted set A_c and use the encoder prediction; the remaining examples form the forwarded set R_c and are either abstained on or passed to one verifier. The gray lane shows the split-conformal alternative to the heuristic confidence gate.

Examples with $\gamma \leq \theta_c$ form the forwarded set R_c. They are either abstained on or passed to one of the candidate verifiers g_c, with $g_c = \varnothing$ corresponding to abstention without verification.

Figure 1 summarizes the resulting routing process. It consists of three decisions. First, the encoder produces a calibrated confidence score. Second, examples whose confidence exceeds the configuration-specific threshold are accepted directly. Third, the remaining examples are either abstained on or forwarded to a verifier.

Verifier space. We consider seven verifier configurations, chosen to test complementary hypotheses about whether a second prediction can improve the uncertain examples: 1) *Zero-shot Qwen3.5-4B*, testing generic compact-LLM capability; 2) *LoRA-tuned Qwen*, testing whether task supervision improves verification; 3) $k = 3$ *voting*, testing whether multiple LLM predictions improve reliability; 4) $k = 5$ *self-consistency*, testing whether sampling variance is a limiting factor; 5) *Heterogeneous-role voting*, testing whether complementary roles provide useful diversity; 6) *MultiAgentESC-style routing*, testing whether explicit multi-agent structure adds value; and 7) *A second encoder*, providing a non-LLM control for whether a second opinion helps independently of LLM capabilities. Two further verifiers enter in Appendix D.7 as robustness checks rather than as search candidates: another LLM (Gemma3-4B) and a size ladder over one generation of Qwen.

Each verifier may also abstain rather than return a label. We additionally include $g_c = \varnothing$ as an eighth, zero-cost option, corresponding to abstention without verification. This asks whether an

additional verifier improves the uncertain examples relative to abstaining.

Cost-quality optimization. Each configuration $c = (\phi, \theta_c, g, T)$ specifies a confidence function, acceptance threshold, verifier slot, and temperature setting (T). We optimize the three-objective vector

$$\mathbf{F}(c) = (F_{1,\text{acc}}(c), \text{Cov}_{5\%}(c), -\text{cost}(c)), \quad (2)$$

where $F_{1,\text{acc}}$ is macro-F1 on the accepted set A_c, $\text{cost}(c)$ is the cost of configuration (see next paragraph), and $\text{Cov}_{5\%}$ is the fraction of examples accepted while satisfying a 5% empirical error constraint on the accepted set. We set $\text{Cov}_{5\%} = 0$ when this constraint is violated.

The coverage constraint prevents a degenerate router from achieving high accepted-set F1 by accepting examples without maintaining a sufficiently low error rate. It therefore focuses the search on configurations that are both accurate and useful at their claimed coverage.

NSGA-II is used to enumerate the Pareto-optimal configurations rather than selecting a single threshold manually, allowing us to compare escalation against abstention under the same cost-quality objective. Thus, the search exposes the full cost-quality frontier.

Cost model. For the search, we use a normalized cost model ($\text{cost}(c)$ in Eq. 2) in which one encoder call costs 1.0, confidence computation costs 0.1, and one compact-LLM call costs 10.0, with each cost weighted by its empirical call rate.

This proxy deliberately *underprices* LLM inference relative to the measured deployment cost. On the machine used for our experiments, one compact-LLM call costs 70-102$\times$ an encoder forward, while a multi-call committee costs 735-1232$\times$ (Table 5 in Appendix D.1). Consequently, the proxy is favorable to LLM verifiers: if they fail to reach the Pareto frontier under the $10\times$ proxy, increasing their cost to the measured ratios cannot make them more competitive.

Importantly, our rejected-slice analysis is independent of this cost model, directly comparing verifier and encoder accuracy on the same forwarded set R_c to separate the selection effect from the inference-cost argument. Remaining hyperparameters are in Appendix C.

Split-conformal alternative. The threshold-based gate provides an empirical accepted-error constraint, but this guarantee need not transfer from validation to test data. We therefore evaluate split-conformal prediction as a principled alternative to the heuristic confidence threshold (Angelopoulos and Bates, 2023).

On a calibration set, we compute the nonconformity scores $s_i = 1 - p_{y_i}$ and let $\hat{q}$ be the $\lceil (n+1)(1-\alpha) \rceil$-th smallest calibration score. For a new example x, the conformal prediction set is

$$S(x) = \{y : p_y \geq 1 - \hat{q}\}. \quad (3)$$

We accept the encoder prediction only when $S(x)$ is a singleton; otherwise, the example is abstained on. This provides a distribution-free alternative to threshold-based selective prediction and tests whether our conclusions depend on the particular choice of confidence threshold.

The methodology separates three questions: which examples should be accepted, whether an additional verifier improves the remaining examples, and whether that improvement justifies its inference cost. Section 3 uses this framework to determine whether compact LLM verification improves the practical Pareto frontier.

3 Experiments

We use three datasets collapsed to Ekman-7 labels: GoEmotions (GE) (Demszky et al., 2020), EmpatheticDialogues (ED) (Rashkin et al., 2019), and Russian CEDR (Sboev et al., 2021). Datasets, label mappings, encoder backbones (RoBERTa-base, XLM-R-base, DeBERTa-v3-base), the LLM baselines and the non-LLM k=3 RoBERTa ensemble, and the evaluation metrics are specified in Appendix C.

Main Results. Table 1 gives the central comparison. On GoEmotions, the encoder-first router reaches $\mathbf{0.828 \pm 0.003}$ accepted-set macro-F1 at $22.6 \pm 0.4\%$ coverage with no LLM calls, while the validation-selected 5% empirical-risk gate lets $\approx 8.7\%$ accepted error through on test, motivating the conformal alternative. Every full-coverage LLM system underperforms the encoder there while costing 70-1232$\times$ more per example (per-system numbers: Appendix D.2), and a k=3 RoBERTa ensemble at $3\times$ cost reaches $F_1 = 0.659$ on its unanimous-accept subset (87.5% coverage): a stronger low-cost non-LLM alternative than any $k \geq 3$ LLM committee we tried.

ED is the boundary of the claim rather than an instance of it: its encoder is the GoEmotions-trained RoBERTa applied zero-shot, scores 0.493 ± 0.004, and is beaten standalone by 7-shot Qwen (0.558).

Table 1: Main quality-cost comparison. *Block 1*: full-coverage systems, which label every example and anchor each component's standalone quality. *Block 2*: selective routers, which may abstain or escalate; each row is one configuration of the search space of Section 2 (full front: Fig. 3 in Appendix D.2). *Block 3*: the split-conformal rule we recommend instead of escalation. *Enc./LLM*: model calls per example. *Rel.lat.*: measured wall-clock per example relative to one encoder forward (Table 5 in Appendix D.1). $F_{1,\text{full}}$/$F_{1,\text{sel}}$: macro-F1 over all/accepted examples. **Bold** marks the best value in each quality column, per dataset; *Coverage* and the cost columns are not marked, because for them more is not better: coverage trades against $F_{1,\text{sel}}$ and is read together with it. $F_{1,\text{full}}$ is left as "–" wherever the row's full-coverage stream is the Block 1 encoder itself: abstention never overrides an encoder prediction, so those rows would repeat one number four times. Cells are mean±std over seeds $\{42, 43, 44\}$ unless marked † (deterministic single-seed LLM baseline); ‡ is the sole LLM-slot Pareto endpoint (seed 43), a maximum-cost endpoint. The Qwen ED result uses a 7-shot prompt. CEDR conformal rows and all three miscoverage levels are in Appendix D.10, Table 18.

	Calls / latency			GoEmotions			EmpatheticDialogues		
System	Enc.	LLM	Rel.lat.	$F_{1,\text{full}}$	$F_{1,\text{sel}}$	Coverage	$F_{1,\text{full}}$	$F_{1,\text{sel}}$	Coverage
Block 1: full-coverage systems (no abstention)									
RoBERTa FT	1	0	1×	**0.623±0.005**	–	–	0.493±0.004	–	–
Qwen3.5-4B (single)†	0	1	102×	0.429	–	–	**0.558**	–	–
multi-call LLM committees†	0	3-5	735-1232×	0.341-0.442	–	–	–	–	–
LoRA Qwen3.5-4B†	0	1	70×	0.608	–	–	0.534	–	–
Block 2: selective routers (may abstain or escalate)									
Encoder-first selective router	1	0	1×	–	**0.828±0.003**	0.226±0.004	–	–	0.000±0.000
DeBERTa-v3-base (GE only)	1	0	1.5×	0.614±0.005	0.760±0.012	0.239±0.007	–	–	–
k=3 RoBERTa unanimous subset	3	0	3×	–	0.659	0.875	–	–	–
Sole LLM-slot Pareto endpoint‡	1	2.09	≈ 516×	0.470	0.824	0.221	–	–	–
Block 3: split-conformal abstention sweep (3 seeds, marginal)									
Conformal (α=0.05)	1	0	1×	–	0.807±0.022	0.242±0.042	–	0.132±0.006	0.010±0.003
Conformal (α=0.10)	1	0	1×	–	0.817±0.005	0.366±0.023	–	0.256±0.017	0.071±0.009
Conformal (α=0.20)	1	0	1×	–	0.705±0.012	0.655±0.019	–	**0.525±0.010**	0.251±0.001

Abstention accordingly has nothing to accept – coverage is 0 under the 5% budget on all three seeds, and conformal needs $\alpha = 0.20$ for $25.1 \pm 0.1\%$ acceptance at $F_{1,\text{acc}} = 0.525 \pm 0.010$ – whereas escalation does pay, making ED the counterexample to our main setting (Table 2). The CEDR and DeBERTa-v3-base rows reproduce the pattern across languages and encoder families (Fig. 2).

Why Compact LLM Verifiers Fail? All seven verifier slots entered the NSGA-II search on equal footing, including LoRA-fine-tuned Qwen (standalone $F_1 = 0.608$). Across all seeds only encoder-only and second-encoder configurations populate the high-quality low-cost region of the front; the single surviving LLM-slot point is a degenerate k=3 voting configuration at $\approx 516\times$ measured latency and full-test $F_1 = 0.470$, a maximum-cost endpoint rather than a frontier improvement. Thus, **no compact-LLM verifier improves the high-quality low-cost Pareto frontier**, and a cost-ratio sweep over $\{2, 5, 10, 20, 50\}\times$ agrees, admitting an LLM configuration only at an unrealistic $2\times$ (front and sweep: Appendix D.2).

The selection effect. Uncertainty gating forwards a conditional slice (the encoder's low-confidence region) on which compact LLMs are no better than, or much worse than, the encoder. This is a distributional mismatch, not a general LLM failure: the examples the encoder is unsure about are not those where an LLM adds information. In Table 7, LoRA-Qwen is indistinguishable from the encoder there ($\Delta = -0.002 \pm 0.002$, Stouffer-combined $p = 0.68$) while still costing an extra call, and every other verifier is significantly worse ($\Delta \in [-0.27, -0.16]$, $p < 0.001$), even though LoRA-Qwen scores 0.608 on the full split, close to encoder parity. On the forwarded slice, this comparison is independent of cost: no compact LLM would help even at zero cost. A second family, Gemma3-4B, behaves identically (Appendix D.7), so the conclusion is not Qwen-specific.

What are the forwarded examples? Joining GE to its multi-rater annotations shows the forwarded slice is more ambiguous (mean annotator plurality 0.713 vs. 0.882, Mann-Whitney $p < 10^{-169}$), more minority-heavy (40.1% vs. 16.4% non-*joy*/*neutral*) and more novel in embedding space (kNN distance 0.535 vs. 0.347, $p \approx 0$). Binned by annotator agreement (Table 2a; figure form in Appendix D.5), encoder and verifier accuracy rise and fall *together*: the same variable governs both, which is why a same-taxonomy verifier has little to recover, and with 70% of the slice at plurality ≤ 0.75 the dominant failure source is

Table 2: The mechanism, and how far it reaches. *(a)* The forwarded slice on GoEmotions, binned by the fraction of annotators who agreed on the gold label. Encoder and verifier accuracy rise and fall *together* with human agreement, so what the gate forwards is not a knowledge gap a same-taxonomy verifier could close; *all* is the paper's operating budget, i.e. the top block of Table 7. *(b)* The identical gated comparison on all four tasks, with the 4B verifier and, indented, the largest rung of the ladder. *Enc.* F_1/*LLM* F_1: standalone macro-F1 on the same evaluation set. Δ: verifier minus encoder accuracy on the slice the gate forwards. *Hybrid*: end-to-end macro-F1 when the forwarded examples are answered by the verifier, to be read against *Enc.* F_1 in the same row block. Every Δ has Stouffer-combined exact McNemar $p < 10^{-3}$ except those marked †, which the paired test cannot separate from zero. All four tasks use the same zero-shot protocol so that they are one experiment rather than four, which is why the ED verifier scores 0.564 here rather than the 0.558 of the 7-shot baseline in Table 1. Budgets of 10, 40 and 80%, the remaining verifiers, and the full ladder behind each gate are in Appendix D.8.

(a) GoEmotions: the forwarded slice, by annotator plurality agreement

Accuracy on the slice	≤ 0.50	0.50-0.75	0.75-1.00	1.00	all
Share of the slice	14.3%	55.7%	5.3%	24.8%	100%
Encoder	0.400	0.549	0.727	0.746	0.586
Qwen3.5-4B LoRA	0.419	0.560	0.677	0.713	0.584
Qwen3.5-4B zero-shot	0.362	0.391	0.473	0.449	0.406

(b) All four tasks, one protocol, two forwarding budgets

			budget 20%		budget 60%	
Task (encoder)	Enc. F_1	LLM F_1	Δ	Hybrid	Δ	Hybrid
GoEmotions (in-domain)	0.618	0.429	−0.088	0.600	−0.167	0.501
Qwen3.7-Max	–	0.392	−0.138	0.566	−0.211	0.460
AG News (in-domain)	0.936	0.812	−0.277	0.879	−0.179	0.824
Qwen3.7-Max	–	0.885	−0.138	0.908	−0.079	0.888
CEDR, ru (in-domain)	0.841	0.749	$-0.073^{\dagger}$	0.846	−0.147	0.795
Qwen3.7-Max	–	0.828	$+0.003^{\dagger}$	0.860	−0.047	0.848
EmpatheticDial. (transferred)	0.493	0.564	+0.142	0.528	+0.086	0.565
Qwen3.7-Max	–	0.650	+0.252	0.559	+0.184	0.636

irreducible label ambiguity rather than a knowledge gap a compact LLM could close. Accordingly LoRA-Qwen's corrections are almost entirely a *neutral* bias cancelled by corruptions of *joy*, *anger*, and *surprise* (Table 11, Appendix D.6).

Neither the trigger nor the verifier's size rescues it. Re-scoring every verifier under six alternative gates, disagreement, a learned router, representation-space novelty (Lakshminarayanan et al., 2017; Lee et al., 2018; Sun et al., 2022), leaves the effect intact, and so does an *oracle* forwarding exactly the encoder's errors, which upper-bounds any routing rule ($+0.003$, $p = 0.38$; Appendices D.3, D.4). Nor is 3-4B the constraint: from 9B and 27B dense through 122B-A10B and 397B-A17B mixture-of-experts (Qwen Team, 2026a) to Qwen3.7-Max (Qwen Team, 2026b), the slice deficit stays flat at $\Delta \in [-0.221, -0.181]$ while end-to-end quality *falls* as the verifier grows ($0.600 \rightarrow 0.566$ at a 20% budget, against 0.618 ± 0.004 encoder-only). Parity still belongs to the 4B model with task supervision, and the flat ladder is specific to this dataset's labelling convention, not a ceiling on the models (Appendix D.7).

Where does escalation pay, then? To test whether the mechanism generalizes, we ran the identical gated comparison (same gate, matched budgets, same paired tests) on all four tasks (Table 2b; full grid in Appendix D.8). What separates them is neither affect nor language, but how far the encoder outruns the verifier standalone, which is set by whether it saw the task's own labels. Where that margin is wide – GoEmotions (0.618 vs. 0.429) and the non-affective AG News (0.936 vs. 0.812) (Zhang et al., 2015) – escalation loses at every budget and every scale. Where it is narrow (Russian CEDR, encoder 0.841 against Qwen3.7-Max's 0.828), the two are indistinguishable at small budgets, and escalation turns into a loss only as the budget grows. And where it runs the other way (EmpatheticDialogues, whose encoder is transferred zero-shot and is beaten standalone), escalation wins outright, by more the larger the verifier ($0.493 \rightarrow 0.636$ at a 60% budget). ED is the positive control: the measurement that finds no benefit on three tasks finds a large one on the fourth, precisely where our precondition of an in-domain encoder fails.

4 Conclusion

Given an in-domain encoder, encoder-first routers and non-LLM alternatives own the high-quality low-cost frontier – across six gates, an oracle, two LLM families, a $\approx 100\times$ scale range, two encoder families, and three tasks in two languages. The mechanism is a selection effect grounded in annotation ambiguity: gates forward examples humans themselves disagree on, encoder and verifier fail there together, and neither scale nor task supervision changes that. The precondition is sharp: on EmpatheticDialogues, whose encoder is transferred, the same measurement finds escalation clearly worth its cost. What decides it is not domain, language, or verifier size, but whether the encoder saw the task's own labels. Given one, compare the verifier on the forwarded slice rather than the full test set, sweep the budget before buying LLM inference, and on a tie prefer conformal abstention or a small ensemble; without one, evaluate the LLM as a full-coverage classifier first.

Limitations

Four boundaries bound what we tested. First, we fine-tuned a verifier only at 4B: a *supervised* verifier above that is untested, and since LoRA at 4B is the one configuration reaching encoder parity, it is the direction the ladder most clearly leaves open. Second, the gated comparison covers all four tasks (Appendix D.8) but the *characterization* of the forwarded slice does not: GoEmotions is the only one released with per-annotator votes, so elsewhere we can show that encoder and verifier fail on the same examples, but not that human annotators disagreed about them. Third, serving precision varies along the ladder – the 4B rungs are locally quantized, the API rungs served at bf16/fp8 – and Qwen3.8-Max could not be evaluated at all, its endpoint refusing to disable reasoning (Appendix D.7). Fourth, latency is measured on one Apple-Silicon/Ollama setup; the non-affective replication uses one dataset and one verifier family, and CEDR uses two seeds over a single XLM-R checkpoint. We make no claim about training-time LLM use, larger encoders, or multimodal benchmarks such as MELD and IEMOCAP (Poria et al., 2019; Busso et al., 2008); our labels are categorical, and carrying the protocol to dimensional ones (Russell, 1980; Mohammad, 2018) through conformal regression (Romano et al., 2019) is future work, as are SelectiveNet-style trained abstention and conformal prediction over LLM posteriors.

Ethics Statement

All datasets are public and were collected by their original authors. Emotion classification has dual-use risks (surveillance, user profiling, psychological inference); we do not recommend this pipeline for clinical, hiring, or other high-stakes decisions without task-specific validation, consent, and human oversight. Because the datasets contain user-generated text, they may include sensitive, offensive, or personally revealing content. We rely on the original dataset collection and anonymization procedures, do not attempt to identify individuals, and do not redistribute raw dataset examples. We use all datasets and checkpoints under their publicly stated licenses or access terms; the released artifact contains code, scripts, configuration files, and derived/cached verifier outputs for reproducing experiments from the original sources. The size-ladder predictions (Appendix D.7) were obtained through a third-party inference API (OpenRouter, routing to the Alibaba and DeepInfra endpoints), so those test examples were transmitted to an external provider; the datasets involved are public, we sent no additional data, and the resulting predictions are cached and released like every other verifier output. AI assistants were used for language polishing, code assistance, and drafting support; the authors reviewed all claims, experiments, analyses, and final text and take full responsibility for the submission.

Acknowledgments

The work of A. Savchenko was implemented in the framework of the Basic Research Program at HSE University (HSE-BR-2025-080).

References

Anastasios N. Angelopoulos and Stephen Bates. 2023. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591. ArXiv:2107.07511.

Hediyeh Baban, Sai A. Pidapar, Aashutosh Nema, and Sichen Lu. 2025. Enhancing text classification with a novel multi-agent collaboration framework leveraging BERT. *arXiv preprint arXiv:2502.18653*.

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 (NeurIPS)*, volume 33, pages 1877–1901.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2024a. AutoAgents: A framework for automatic agent generation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 22–30.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024b. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research (TMLR)*. ArXiv:2305.05176.

C. K. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Polina Demochkina and Andrey V Savchenko. 2021. MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices. In *Proceedings of International Conference on Pattern Recognition (ICPR) Workshops*, pages 266–274. Springer.

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 (ACL)*, pages 4040–4054.

Paul Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In James Cole, editor, *Nebraska Symposium on Motivation, 1971*, volume 19, pages 207–282. University of Nebraska Press, Lincoln.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.

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

Gemma Team. 2025. Gemma 3 technical report. arXiv:2503.19786.

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)*, pages 1321–1330.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. ArXiv:2111.09543.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. ArXiv:2106.09685.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. ArXiv:1612.01474.

Danila Lapokin and Andrey Savchenko. 2026. CRL-Prompt: Contrastive and reinforcement learning for soft prompt tuning for text classification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1391–1398.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. ArXiv:1807.03888.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

David Madras, Toniann Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. ArXiv:1711.06664.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 174–184.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 7076–7087. ArXiv:2006.01862.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, and 1 others. 2024a. EAI: Emotional decision-making of llms in strategic games and ethical dilemmas. *Advances in Neural Information Processing Systems*, 37:53969–54002.

Mikhail Mozikov, Nikita Severin, Maria Glushanina, Mikhail Baklashkin, Andrey Savchenko, and Ilya Makarov. 2024b. InsideOut: Unifying emotional llms to foster empathy. In *Proceedings of 27th European Conference on Artificial Intelligence (ECAI)*, pages 4499–4502.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2901–2907.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to route LLMs from preference data. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10671–10682.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Blog post.

Qwen Team. 2026b. Qwen3.7: The agent frontier. https://qwen.ai/blog?id=qwen3.7. Blog post; Qwen3.7-Max is closed-weights and API-only.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5370–5381.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. 2019. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. ArXiv:1905.03222.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. 2020. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33. ArXiv:2006.02544.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Andrey V Savchenko and Lyudmila V Savchenko. 2024. Inside Out: Emotional multiagent multimodal dialogue systems. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), demo track*, pages 8784–8788.

Alexander Sboev, Aleksandr Naumov, and Roman Rybka. 2021. Data-driven model for emotion detection in Russian texts. *Procedia Computer Science*, 190:637–642.

Jinwei Su, Qizhen Lan, Yinghui Xia, Lifan Sun, Weiyou Tian, Tianyu Shi, Xinyuan Song, Lewei He, and Jingsong Yang. 2026. Difficulty-aware agentic orchestration for query-specific multi-agent workflows. In *Proceedings of the ACM Web Conference 2026 (WWW)*. ArXiv:2509.11079.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 20827–20840. ArXiv:2204.06507.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Neeraj Varshney and Chitta Baral. 2022. Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11007–11021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. ArXiv:2203.11171.

Jiawei Xu, Arief Koesdwiady, Sisong Bei, Yan Han, Baixiang Huang, Dakuo Wang, and 1 others. 2026. Rethinking the value of multi-agent workflow: A strong single agent baseline. *arXiv preprint arXiv:2601.12307*.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4665–4681.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. 2025. EvoAgent: Towards automatic multi-agent generation via evolutionary algorithms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1: Long Papers*, pages 6192–6217. ArXiv:2406.14228.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. 2025a. Multi-agent architecture search via agentic supernet. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. ArXiv:2502.04180, Oral presentation.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. ArXiv:1509.01626.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025b. DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, page 107901.

A Notation

Table 3 collects all symbols introduced in Section 2 and their meanings.

Table 3: Notation.

Symbol	Meaning
x, y	input text and its gold label
f_θ	fine-tuned encoder; $\ell = f_\theta(x)$ logits
K	number of closed labels (7 for Ekman)
T	temperature fit on validation data
p	$\mathrm{softmax}(\ell/T)$, calibrated posterior
$\hat{y}$	$\arg\max_y p_y$, the encoder label
ϕ	confidence function (margin, max prob., entropy)
γ	$\phi(p) \in [0, 1]$, the gate score
θ_c	acceptance threshold of configuration c
g	verifier slot ($\varnothing$ if unused)
$\perp$	abstention
α	target conformal miscoverage
$\hat{q}$	conformal quantile of calibration scores
$S(x)$	conformal prediction set
budget	fraction of examples forwarded by a gate

B Related Work

Our work connects three lines of research, and we state our position with respect to each.

Multi-agent systems and agent search. AutoAgents, EvoAgent, and MaAS search over LLM-agent roles and topologies (Chen et al., 2024a; Yuan et al., 2025; Zhang et al., 2025a), and recent analyses question whether such graphs justify their overhead (Xu et al., 2026). Emotion-oriented MAS systems (Mozikov et al., 2024b; Savchenko and Savchenko, 2024) such as BertMAS and MultiAgentESC motivate our multi-call baselines (Baban et al., 2025; Xu et al., 2025); a scoping run of AutoAgents and EvoAgent on our tasks places them outside the low-overhead regime (Appendix D.13). These works add agents to improve quality; we take the complementary decision-theoretic view (given an existing supervised encoder, when is adding a compact LLM agent Pareto-justified?) and give the first controlled negative answer for closed-label emotion.

Selective prediction, deferral, and conformal inference. Rejection has been studied since Chow (1970), through selective classifiers (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017), calibration (Guo et al., 2017), learning to defer (Madras et al., 2018; Mozannar and Sontag, 2020), model cascades (Varshney and Baral, 2022), and conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2023; Romano et al., 2020). We do not extend this theory but import it, placing abstention *in direct competition* with LLM escalation inside one cost-aware search: when abstention dominates, do not escalate. **LLM routing** work such as FrugalGPT, RouteLLM, and DAAO studies cost-aware selection among generative LLMs that all see the same input distribution (Chen et al., 2024b; Ong et al., 2025; Su et al., 2026); we route from a supervised discriminative encoder, so the forwarded set is an uncertainty-conditioned subsample rather than a random query, and on it the compact LLM's advantage disappears – a selection effect not, to our knowledge, documented in that literature. Prior work on GoEmotions (Demszky et al., 2020) and EmpatheticDialogues (Rashkin et al., 2019) relies on supervised encoders (Liu et al., 2019; He et al., 2023); we fix a base-size encoder to isolate the routing question.

C Experimental Setup

C.1 Datasets and Metrics

We use three datasets collapsed to Ekman-7 labels: 1) **GoEmotions (GE)** (Demszky et al., 2020): 28 original labels (27 emotions plus neutral) $\rightarrow$ 6 Ekman emotions + neutral ($K = 7$); train/val/test: 43,410/5,426/5,427; 2) **EmpatheticDialogues (ED)** (Rashkin et al., 2019): 2,542 test utterances; 32 situation labels $\rightarrow$ 6 Ekman emotions + neutral. The encoder is the GE-trained RoBERTa applied zero-shot (transfer setting); and 3) **Russian CEDR** (Sboev et al., 2021): 4,037 train, used to fine-tune an XLM-R-base encoder; mapped to 5 of 7 Ekman labels (no disgust, no separate neutral), so we treat it as a non-English diagnostic rather than a directly comparable main benchmark.

Calibration sets are stratified subsamples ($n = 500/400/400$): GE and ED use validation splits; CEDR partitions its test split into disjoint calibration and evaluation subsets because no validation split is available. Calibration examples are never included in reported evaluation metrics.

Macro-F1 on full test split and on accepted subset, acceptance rate (coverage), per-sample model calls, and relative latency (encoder forward = $1\times$).

C.2 Label Mappings

For **GoEmotions** (Demszky et al., 2020), we apply the official Ekman grouping released by the dataset authors (Demszky et al., 2020) without modification: *anger* $\leftarrow$ anger, annoyance, disapproval; *disgust* $\leftarrow$ disgust; *fear* $\leftarrow$ fear, nervousness; *joy* $\leftarrow$ joy, amusement, approval, excitement, gratitude,

love, optimism, relief, pride, admiration, desire, caring; *sadness* ← sadness, disappointment, embarrassment, grief, remorse; *surprise* ← surprise, realization, confusion, curiosity; *neutral* ← neutral.

For **EmpatheticDialogues** (Rashkin et al., 2019), the 32 situation labels are mapped by a fixed dictionary released with our code, following the lexical correspondence of the original label names: *anger* ← angry, furious, annoyed; *disgust* ← disgusted; *fear* ← afraid, terrified, anxious, apprehensive; *joy* ← joyful, excited, proud, grateful, hopeful, trusting, content, anticipating, confident, caring, faithful; *sadness* ← sad, devastated, guilty, ashamed, lonely, jealous, embarrassed, disappointed; *surprise* ← surprised, impressed; *neutral* ← sentimental, nostalgic, prepared. The mapping is fixed before any experiment and is identical for the encoder and every LLM baseline, so it cannot favour either system.

CEDR (Sboev et al., 2021) is already annotated in an Ekman subset (joy, sadness, surprise, fear, anger) and is used unchanged; it has no *disgust* and no separate *neutral* class, which is why we treat it as a diagnostic rather than a directly comparable benchmark. **AG News** uses its four original topic labels (world, sports, business, sci/tech) with no mapping.

C.3 Models and Baselines

We use the following **Encoders**: RoBERTa-base (Liu et al., 2019) fine-tuned on GE (3 epochs, batch 32, lr 2×10^{-5}, seeds 42,43,44); XLM-R-base (Conneau et al., 2020) on CEDR; DeBERTa-v3-base (He et al., 2023) as a robustness check. The **LLM backbone** is Qwen3.5-4B (Qwen Team, 2026a) (Ollama, greedy decoding, temperature 0), with zero-shot Gemma3-4B (Gemma Team, 2025) as a second family (Appendix D.7). Unless noted otherwise, LLM baselines are deterministic (greedy), so variability across seeds comes only from the encoder; this isolates the routing question from sampling noise. We have several **baselines**: single Qwen (zero-shot on GE; 7-shot (Brown et al., 2020) on ED - asymmetry favours LLM on ED, making our conclusion conservative), homogeneous voting k=3, self-consistency k=5 (Wang et al., 2023), heterogeneous voting (Lexical/Contextual/Logic roles, k=3) inspired by BertMAS (Baban et al., 2025), MultiAgentESC-style pipeline (Xu et al., 2025), and LoRA-fine-tuned (Hu et al., 2022) Qwen (rank 16, $\alpha = 32$, trained on GE).

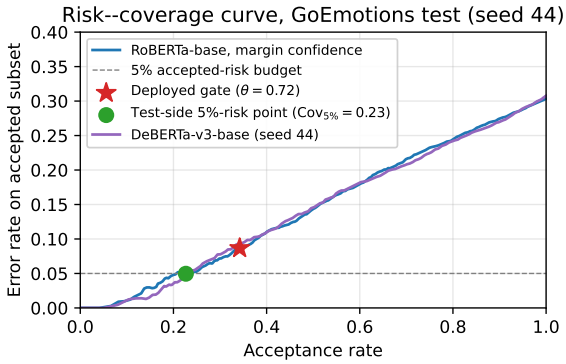


Figure 2: Risk-coverage curve on GoEmotions test (seed 44, margin confidence). The test-side 5%-risk operating point accepts $\approx 22.7\%$ of examples, close to the validation-selected coverage (22.6% in Table 1). DeBERTa-v3-base curve nearly coincides.

In addition, we consider a **non-LLM cost baseline**: a k=3 RoBERTa ensemble across seeds, with unanimous agreement used as a selective-acceptance rule.

C.4 Search Hyperparameters

NSGA-II runs for 15 generations with a population of 20 (300 configuration evaluations per seed) over the configuration space of Section 2 (confidence function, threshold $\theta_c \in [0.30, 0.99]$, verifier slot, temperature switch), evaluating each configuration on the validation split and reporting test metrics for the selected configurations. Verifier outputs are cached per example, so the search prices a verifier slot without re-running inference.

C.5 LoRA Fine-Tuning Details

LoRA on Qwen3.5-4B was fitted on GoEmotions with rank $r = 16$, $\alpha = 32$, dropout 0.05, target modules $\{q_proj, v_proj\}$, learning rate 2×10^{-4}, 3 epochs, batch size 4, gradient accumulation 8 (effective batch 32). Training used TRL's SFTTrainer on the GoEmotions train split, seed 42, so LoRA is a deterministic single-seed baseline; multi-seed uncertainty in Table 7 comes from the encoder seeds and their rejected slices. The resulting model reaches near-parity with RoBERTa in-domain ($F_1 = 0.608$ vs. 0.623 ± 0.005) but drops to 0.534 on EmpatheticDialogues, -0.024 below the 7-shot Qwen baseline. Per-class breakdown shows the *neutral* class dropping from $F_1 = 0.66$ on GE to 0.20 on ED, indicating that LoRA encodes a domain-specific decision boundary for "neutral vs. emotional" that does not transfer.

C.6 Encoder Calibration Diagnostics

Table 4 reports calibration of the RoBERTa encoder on GE (in-domain) and ED (zero-shot transfer). The fitted temperature is stable across seeds: $\hat{T} = 1.30 \pm 0.02$ on GE and $\hat{T} = 2.00 \pm 0.03$ on ED (mean±std, seeds $\{42, 43, 44\}$). In-domain ECE (Naeini et al., 2015) is low (0.020, seed 42); under domain shift ECE more than triples (0.140, seed 42), indicating overconfidence. This explains why the conformal wrapper on ED requires larger α to achieve non-trivial acceptance. For CEDR, the in-domain XLM-R encoder calibrates to $\hat{T} = 1.24$ (seed 42), similar to the GE case.

Table 4: Encoder calibration on the calibration split before applying temperature scaling; $\hat{T}$ is then used for inference-time probabilities. $\hat{T}$ reports mean±std over 3 seeds for GE and ED; ECE and Brier are seed-42 only (refreshing them across seeds requires re-running per-seed calibration with logits export).

Dataset	Encoder	ECE	Brier	$\hat{T}$	Accuracy
GE	RoBERTa (in-domain)	0.020	0.430	1.30±0.02	0.687
ED	RoBERTa (zero-shot)	0.140	0.620	2.00±0.03	0.552

D Additional Results

Everything below supports a claim made in Section 3: the cost measurements behind the frontier, the full grids the main tables summarize, the analyses that identify the mechanism, and the controls that test how far it reaches. Each subsection names the script and artifact that regenerate it in our source code.

D.1 Measured Deployment Latency

Table 5 reports wall-clock latency measured on the machine that produced every result in this paper (Apple Silicon; encoders on MPS, LLMs served by Ollama), averaged over the full GoEmotions test split. The 10× proxy used inside the search understates the measured gap by one to two orders of magnitude, so the search was biased in favour of LLM verifiers and the negative result holds a fortiori. We do not have energy instrumentation; on fixed hardware, wall-clock and call counts are monotone proxies for it. Latency is reported for sequential execution, which is how the router would run a single request; the parenthesized wall-clock column reflects the concurrency used to produce the cached predictions.

The table lists only verifiers we ran locally. The size-ladder rungs of Appendix D.7 are served by a third-party API, so their wall-clock is dominated by the network and by another tenant's batching and is not comparable to these measurements; for them we report the call count, which is. Their deployment cost is nonetheless strictly worse than anything in the table: none of them runs on the hardware assumed throughout this paper, since 122B-A10B and 397B-A17B do not fit and Qwen3.7-Max has no public weights at all. That strengthens the cost argument rather than weakening it.

Table 5: Measured per-example latency on GoEmotions for every verifier that runs on the hardware assumed in this paper. Ratio is relative to one RoBERTa-base forward at batch size 1 (0.0072 s; 0.0051 s at batch size 32).

System	LLM calls	s/example	Ratio
RoBERTa-base (encoder)	0	0.0072	1×
LoRA Qwen3.5-4B	1	0.503	70×
Zero-shot Qwen3.5-4B	1	0.740	102×
Zero-shot Gemma3-4B	1	1.533	212×
Heterogeneous voting k=3	3	5.315	735×
Voting k=3	3	5.339	738×
MultiAgentESC	3	7.117	984×
Self-consistency k=5	5	8.911	1232×

D.2 The Verifier-Slot Pareto Front

Fig. 3 shows the front the search of Section 2 returns on GoEmotions, aggregated over the three encoder seeds and colored by which verifier occupies each Pareto point's slot. Only encoder-only and second-encoder configurations reach the high-quality low-cost region. The one surviving LLM-slot point is a degenerate k=3 voting configuration at $\approx 22\times$ normalized search cost and $F_1 = 0.470$, which NSGA-II keeps as the maximum-cost endpoint of the front rather than as a cost-quality improvement.

How much cheaper would an LLM call have to be? The search prices an LLM call at 10× an encoder forward. Re-running it at ratios $\{2, 5, 10, 20, 50\}\times$ gives LLM-slot shares of the front of $\{0.133, 0, 0, 0, 0\}$: only at an unrealistically optimistic 2× does any LLM configuration enter the front at all, and even there under 14% of Pareto configurations use one. The measured ratio is 70-1232× (Appendix D.1), so the search as run was already generous to the verifier.

The multi-call committees, individually. Block 1 of Table 1 collapses the four multi-call LLM systems into one row, because their point is collective: every one of them costs three to five LLM calls per example and none reaches the en-

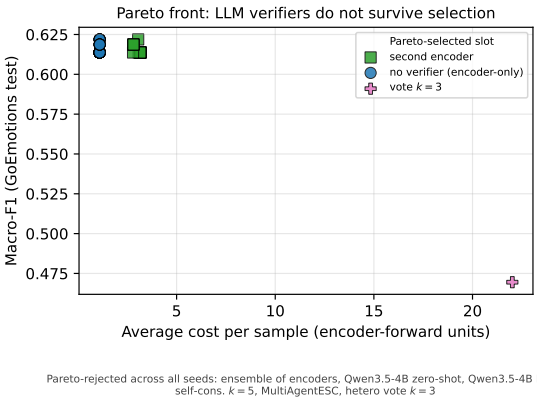


Figure 3: Pareto front of the NSGA-II verifier-slot search on GoEmotions, aggregated over seeds $\{42, 43, 44\}$, colored by the verifier slot of each Pareto point. Only encoder-only and second-encoder configurations populate the high-quality low-cost region. The single k=3 voting point at $\approx 22\times$ normalized search cost / $F_1 = 0.47$ is the maximum-cost extreme that NSGA-II retains as a Pareto endpoint; LoRA-Qwen, zero-shot Qwen, self-consistency, MultiAgentESC, and heterogeneous voting do not enter the high-quality low-cost region of the front.

coder. Table 6 separates them.

Table 6: Full-coverage macro-F1 of the multi-call LLM committees on GoEmotions, the four systems summarized as one row in Block 1 of Table 1. *LLM*: calls per example; *Rel.lat.*: measured wall-clock relative to one encoder forward (Appendix D.1). All are deterministic single-seed baselines, and all fall below the 0.623 ± 0.005 of the encoder they were meant to improve on.

System	LLM	Rel.lat.	F_1
Heterog. voting k=3	3	735×	0.442
Voting k=3	3	738×	0.416
Self-cons. k=5	5	1232×	0.415
MultiAgentESC	3	984×	0.341

Conformal as a search objective. A sceptical reading of the front is that the heuristic gate, not the verifier, drives the rejection. As a methodological cross-check we re-ran the search with split-conformal singleton acceptance ($\alpha = 0.10$) as the acceptance rule, integrated directly into the fitness evaluation. The front stays qualitatively identical to Fig. 3: no LLM verifier improves the high-quality low-cost region under a principled abstention rule either.

D.3 Verifier and Gate Analysis on GoEmotions

Table 7 is the full version of the comparison that was summarized in Section 3: every verifier on the slice the confidence gate forwards, every rung of the size ladder under the same protocol, and the best verifier under every alternative gate at a matched budget.

Does the escalation trigger matter? Every result in the main text uses a confidence gate, so that those results might describe that trigger rather than inference-time verification as such. We therefore re-scored all seven cached verifiers under the six gates listed in Table 7, which span ensemble and cross-architecture disagreement, learned routing (test AUC 0.76), and representation-space novelty (Lakshminarayanan et al., 2017; Lee et al., 2018; Sun et al., 2022), plus an *oracle* forwarding exactly the encoder's errors, which upper-bounds any routing rule. Gates are compared at matched budgets over 3 seeds with the paired tests of Table 7; definitions and the full grid are in Appendix D.4.

The lower panel of Table 7 reports the paper's operating budget (67.9% forwarded). The forwarded subsets genuinely differ – Jaccard overlap with the confidence gate falls to 0.73 here and 0.20 at a 20% budget – yet the selection effect survives every gate, and even a router forwarding precisely the encoder's mistakes extracts no significant gain from the best compact LLM ($+0.003$, $p = 0.38$): at this budget the binding constraint is the verifier's conditional accuracy, not the trigger.

Smaller budgets expose a second constraint: the disagreement and learned gates do give LoRA-Qwen a significant slice gain at a 20% budget ($+0.030$ to $+0.044$), yet end-to-end macro-F1 stays within seed noise of encoder-only (0.616-0.624 vs. 0.618 ± 0.004) at $\approx 15\times$ the latency, while the oracle gate would reach 0.694. Escalation headroom is therefore bounded by *routing* quality at small budgets and by *verifier* quality at large ones.

Cross-checks. Two further checks leave the conclusion unchanged: varying the confidence score (Appendix D.11) and replacing the heuristic gate with the conformal rule inside the search itself (Appendix D.2); the paired rejected-slice tests behind every Δ and p in Table 7 are detailed in Appendix D.12.

Does verifier scale matter? Every verifier so far is 3–4B, so the failure could be a capacity

Table 7: Verifier and gate analysis on GE (3 seeds). Δ is verifier minus encoder accuracy on the forwarded slice, p is Stouffer-combined exact McNemar. *Top*: verifiers behind the confidence gate (encoder accuracy 0.586 ± 0.014); LoRA-Qwen ties, all others are worse. Indented rows are the size ladder – the same zero-shot protocol with only the verifier's scale changed, from 4B to a closed frontier model; the deficit is flat, and Appendix D.7 shows the same ladder rising on three other tasks. *Bottom*: LoRA-Qwen under alternative gates at a matched budget, J = Jaccard overlap of the forwarded set with the confidence gate; the tie survives every gate, including the oracle, and zero-shot Qwen stays worse under all ($\Delta \leq -0.155$, Table 8 in Appendix D.4).

Verifiers behind the confidence gate Verifier	Acc. on slice	Δ vs. enc.	p
Qwen3.5-4B LoRA	0.584 ± 0.012	-0.002 ± 0.002	0.68
Qwen3.5-4B zero-shot	0.406 ± 0.004	-0.181 ± 0.010	$<10^{-3}$
Qwen3.5-9B	0.374 ± 0.008	-0.212 ± 0.006	$<10^{-3}$
Qwen3.5-27B	0.394 ± 0.006	-0.192 ± 0.009	$<10^{-3}$
Qwen3.5-122B-A10B	0.382 ± 0.008	-0.204 ± 0.006	$<10^{-3}$
Qwen3.5-397B-A17B	0.374 ± 0.008	-0.212 ± 0.006	$<10^{-3}$
Qwen3.7-Max	0.365 ± 0.006	-0.221 ± 0.008	$<10^{-3}$
Voting $k=3$	0.418 ± 0.003	-0.168 ± 0.011	$<10^{-3}$
Self-cons. $k=5$	0.413 ± 0.004	-0.174 ± 0.011	$<10^{-3}$
Heterog. voting $k=3$	0.422 ± 0.005	-0.164 ± 0.009	$<10^{-3}$
MultiAgentESC	0.314 ± 0.005	-0.272 ± 0.009	$<10^{-3}$
Gemma3-4B zero-shot	0.326 ± 0.004	-0.261 ± 0.010	$<10^{-3}$
Escalation gates, best verifier (LoRA-Qwen) Gate	J vs. conf.	Δ vs. enc.	p
Confidence (margin)	1.00	-0.002	0.64
Ensemble / cross-arch. disagr.	1.00	-0.002	0.62-0.65
Learned router	0.93	-0.004	0.49
kNN novelty	0.76	-0.005	0.35
Mahalanobis novelty	0.73	-0.002	0.69
Oracle (errors only)	0.93	$+\mathbf{0.003}$	0.38

limit rather than a property of gating. We therefore ran a size ladder inside one model generation under the identical protocol, varying only scale: Qwen3.5 at 9B and 27B (dense) and 122B-A10B and 397B-A17B (mixture-of-experts) (Qwen Team, 2026a), plus the closed frontier model Qwen3.7-Max (Qwen Team, 2026b). The deficit does not shrink. Over a $\approx$100$\times$ range in parameters the slice gap stays flat at $\Delta \in [-0.221, -0.181]$ ($p < 10^{-180}$ at every rung, Table 7), and end-to-end hybrid macro-F1 *decreases* monotonically with size ($0.600 \rightarrow 0.566$ at a 20% budget, against 0.618 ± 0.004 encoder-only). Parity still belongs to the 4B model with task supervision – not a capability ceiling, since Appendix D.7 shows the same ladder rising on three other tasks, so what scale fails to supply here is something other than capacity.

D.4 Alternative Escalation Gates

Gate definitions. All gates score every test example and forward the top-B fraction under a matched budget B, so quality differences cannot come from forwarding different numbers of examples. *Confidence* uses the temperature-scaled margin of Section 2. *Ensemble disagreement* counts argmax disagreements across the three RoBERTa seeds, breaking ties by margin (Lakshminarayanan et al., 2017). *Cross-architecture disagreement* does the same between RoBERTa and DeBERTa-v3-base. The *learned router* is a logistic regression on calibrated confidence features fit on validation data to predict encoder error (test AUC 0.763-0.765 across seeds). *kNN novelty* is the mean L2 distance to the k=10 nearest training embeddings (Sun et al., 2022); *Mahalanobis* is the distance to the nearest class centroid under a Ledoit-Wolf covariance estimated on 10k training examples (Lee et al., 2018). The *oracle* forwards exactly the examples the encoder gets wrong (filling the remaining budget with the lowest-margin correct ones), and therefore upper-bounds any escalation rule at that budget.

Results. Table 8 gives the full grid. Two regimes are visible. At budgets at or above 60% the verifier is the binding constraint: even the oracle gate yields no significant gain for the strongest compact LLM. At a 20% budget the gate becomes the binding constraint: the oracle would lift end-to-end macro-F1 to 0.694, but the best implementable gate reaches 0.624, statistically indistinguishable from encoder-only (0.618 ± 0.004) and from cheaper non-LLM alternatives – a second encoder (0.623 at $\approx 1.3\times$ latency) and the k=3 encoder ensemble (0.622 at $3\times$) – while the LLM hybrid costs $\approx 15\times$. Zero-shot Qwen is significantly worse than the encoder under every gate at every budget.

Note on the encoder-only reference. The gate and slice analyses in this appendix and in Appendix D.5 are computed on the released per-example outputs of the verifier-slot search runs, whose encoder-only full-test macro-F1 is 0.618 ± 0.004; Table 1 aggregates the primary search runs and reports 0.623 ± 0.005. The gap is within seed noise, and every comparison we draw is made within one set of runs, so no conclusion depends on it.

D.5 Characterizing the Forwarded Slice

We joined the GoEmotions test split to the raw multi-rater release by example id (join coverage $5427/5427$), which gives, for every example, the individual annotator votes behind the aggregated gold label. Table 9 compares the forwarded and accepted slices on annotation and representation statistics, and Table 10 bins the forwarded slice by annotator plurality (the figure form is Fig. 4). The forwarded slice is not merely "harder"; it is

Table 8: All escalation gates at matched forwarding budgets (GE, 3 encoder seeds). J: Jaccard overlap of the forwarded set with the confidence gate at the same budget. Δ: verifier minus encoder accuracy on the forwarded set; p is Stouffer-combined exact McNemar for LoRA-Qwen. F_1: end-to-end macro-F1 of the LoRA-Qwen hybrid over the full test split; encoder-only is 0.618 ± 0.004 on the same cached outputs. The last two columns give Δ for zero-shot Qwen and for a DeBERTa second encoder.

		LoRA Qwen			Qwen ZS	2nd enc.
Gate	J	Δ	p	F_1	Δ	Δ
Forwarding budget 20%						
Confidence (margin)	1.00	+0.034	0.002	0.620	−0.088	+0.021
Ensemble disagreement	0.75	+0.044	$<10^{-3}$	0.621	−0.083	+0.020
Cross-arch. disagreement	0.60	+0.044	$<10^{-3}$	0.624	−0.078	−0.014
Learned router	0.56	+0.030	0.010	0.616	−0.085	+0.022
kNN novelty	0.23	+0.010	0.30	0.616	−0.164	−0.004
Mahalanobis novelty	0.20	+0.017	0.06	0.619	−0.152	−0.009
Oracle	0.38	+0.344	$<10^{-3}$	0.694	+0.327	+0.235
Forwarding budget 40%						
Confidence (margin)	1.00	+0.012	0.12	0.615	−0.128	+0.002
Ensemble disagreement	0.98	+0.012	0.10	0.615	−0.128	+0.003
Cross-arch. disagreement	0.93	+0.012	0.11	0.618	−0.129	−0.007
Learned router	0.81	+0.011	0.15	0.614	−0.134	+0.004
kNN novelty	0.43	+0.005	0.52	0.619	−0.173	−0.006
Mahalanobis novelty	0.40	+0.008	0.21	0.618	−0.168	−0.001
Oracle	0.57	+0.127	$<10^{-3}$	0.668	+0.103	+0.051
Forwarding budget 60%						
Confidence (margin)	1.00	+0.002	0.80	0.611	−0.167	−0.003
Ensemble disagreement	1.00	+0.002	0.80	0.611	−0.167	−0.003
Cross-arch. disagreement	0.99	+0.002	0.74	0.612	−0.167	−0.005
Learned router	0.91	−0.000	0.97	0.610	−0.166	−0.003
kNN novelty	0.66	−0.003	0.63	0.611	−0.185	−0.005
Mahalanobis novelty	0.63	−0.003	0.55	0.610	−0.179	−0.004
Oracle	0.87	+0.017	0.002	0.621	−0.112	+0.000
Forwarding budget 67.9% (paper)						
Confidence (margin)	1.00	−0.002	0.64	0.610	−0.181	−0.004
Ensemble disagreement	1.00	−0.002	0.65	0.610	−0.181	−0.004
Cross-arch. disagreement	1.00	−0.002	0.62	0.610	−0.181	−0.004
Learned router	0.93	−0.004	0.49	0.610	−0.182	−0.003
kNN novelty	0.76	−0.005	0.35	0.610	−0.192	−0.004
Mahalanobis novelty	0.73	−0.002	0.69	0.611	−0.189	−0.004
Oracle	0.93	+0.003	0.38	0.614	−0.155	−0.003
Forwarding budget 80%						
Confidence (margin)	1.00	−0.005	0.28	0.608	−0.198	−0.004
Ensemble disagreement	1.00	−0.005	0.28	0.608	−0.198	−0.004
Cross-arch. disagreement	1.00	−0.005	0.28	0.608	−0.198	−0.004
Learned router	0.98	−0.005	0.27	0.608	−0.199	−0.003
kNN novelty	0.87	−0.004	0.34	0.610	−0.203	−0.003
Mahalanobis novelty	0.87	−0.004	0.34	0.609	−0.204	−0.003
Oracle	0.98	−0.004	0.31	0.608	−0.192	−0.004

systematically the region where the human annotators themselves disagreed, and where the gold label is supported by a thin plurality. This is why a verifier that shares the taxonomy – whether a compact LLM or a second encoder – cannot recover it, and it predicts that the residual error is largely irreducible without changing the label definition or the annotation protocol (Plank, 2022; Uma et al., 2021).

D.6 Per-Class Analysis

Table 11 answers which emotions the gate keeps and what the verifier does to them. *Joy* is the only high-confidence class (38.7% forwarded, mean confidence 0.756); every other class is forwarded at 74-88%. On the forwarded slice, LoRA-Qwen's corrections are almost entirely a *neutral* bias (+315 net), which is cancelled by corruptions of *joy*, *anger*, and *surprise* (−308 combined), leaving a net of −8 examples – the aggregate statistical tie of Table 7, now decomposed. Zero-shot Qwen shows the same bias without the compensating accuracy: 483 corrections against 1150 corruptions. The dominant confusions on the forwarded slice are *neutral*↔*joy* and *neutral*↔*anger* for both the encoder and the verifiers, i.e. the two systems fail on the same boundary.

Table 9: Forwarded vs. accepted examples on GoEmotions (mean over 3 encoder seeds). Significance is Mann-Whitney over the pooled per-example values.

Property	Forwarded	Accepted	p
Examples (n)	3685	1742	–
Annotator plurality agreement	0.713	0.882	$<10^{-169}$
Full-agreement rate	24.8%	64.4%	–
Raters supporting the gold label	0.706	0.878	–
Minority-class share	40.1%	16.4%	–
Embedding novelty (kNN dist.)	0.535	0.347	≈ 0

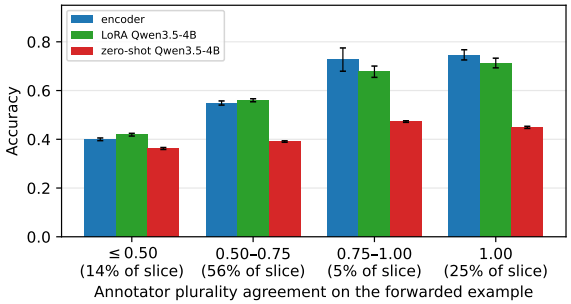


Figure 4: Accuracy on the forwarded slice by annotator plurality agreement (GE, mean±std over 3 encoder seeds). Encoder and verifier accuracy are governed by the same variable – how much the human annotators agreed – which is why a verifier sharing the taxonomy cannot recover the slice.

D.7 Verifier Family and Scale

A second family. To check that the conclusion is not an artifact of the Qwen family, we ran zero-shot Gemma3-4B (Gemma Team, 2025) through the identical protocol (Ollama, greedy decoding, same prompt template, full GoEmotions test split, no parse failures). Standalone macro-F1 is 0.353, below zero-shot Qwen (0.429), and on the rejected slice it is far worse than the encoder ($\Delta = -0.261 \pm 0.010$ over 3 encoder seeds, Stouffer-combined $p \approx 0$). Its measured latency is 212× an encoder forward (Table 5). The selection effect therefore reproduces on a second, independently trained compact-LLM family.

A size ladder. Family robustness leaves a sharper question open: our largest verifier was 4B, so the failure could be a capacity limit rather than a prop-

Table 10: The forwarded slice binned by annotator plurality agreement (GE, mean over 3 seeds). Encoder and verifier accuracy move together with human agreement; 70% of the slice lies in the two lowest-agreement bins.

Plurality	Share	Encoder	LoRA Qwen	Qwen ZS
≤ 0.50	14.3%	0.400	0.419	0.362
0.50-0.75	55.7%	0.549	0.560	0.391
0.75-1.00	5.3%	0.727	0.677	0.473
1.00	24.8%	0.746	0.713	0.449

Table 11: Per-class behaviour on GoEmotions (mean over 3 encoder seeds). Fwd.: fraction of the class forwarded by the gate; Conf.: mean encoder confidence; Enc.: encoder accuracy on the forwarded part of the class; the last three columns count how LoRA-Qwen changes the encoder's answers there.

Class	Fwd.	Conf.	Enc.	n	corr.	corrupt.	net
neutral	87.9%	0.485	0.577	1412	345	30	+315
sadness	74.5%	0.537	0.436	236	24	24	−0
fear	74.1%	0.550	0.622	67	4	11	−7
disgust	85.3%	0.496	0.394	72	2	9	−7
surprise	87.7%	0.490	0.618	502	21	96	−75
anger	85.4%	0.497	0.558	600	35	115	−79
joy	38.7%	0.756	0.662	795	32	186	−154
Total	67.9%	–	0.586	3685	463	471	−8

erty of gating. We therefore fixed everything except scale. The prompt, the verbalizer, greedy decoding (T=0, 32 output tokens), the parser and the full 5427-example test split are shared with the 4B runs, and reasoning is disabled at every rung (no rung billed a single reasoning token). The criterion for a rung to *win* – a significant positive slice Δ (Stouffer $p < 0.05$ over the three encoder seeds) *and* an end-to-end hybrid macro-F1 above encoder-only beyond seed noise – was fixed before any rung ran, and every rung we ran is reported.

The ladder stays inside the Qwen3.5 generation that produced the paper's 4B backbone, so the 4B result lies on the same curve rather than beside it: 9B and 27B dense, then 122B-A10B and 397B-A17B mixture-of-experts (Qwen Team, 2026a), all open-weight, plus the closed frontier model Qwen3.7-Max (Qwen Team, 2026b) as an upper bound from what is available. Table 12 and Fig. 5 report the result: no rung wins, and none comes close. Across a $\approx$100$\times$ span of parameters the slice deficit is flat, standalone macro-F1 varies by less than 0.04, and hybrid macro-F1 *falls* as the verifier grows – at a 20% budget from 0.600 (4B) to 0.566 (Qwen3.7-Max), every value below the 0.618$\pm$0.004 of encoder-only.

Is this a capability ceiling? A flat ladder could mean the models cannot do the task. To separate

Table 12: Verifier size ladder on GE, zero-shot, one protocol, 3 encoder seeds. *Act.* is active parameters (= total for dense models). Δ is verifier minus encoder accuracy on the forwarded slice (encoder 0.586$\pm$0.014); every Δ has Stouffer-combined exact McNemar $p < 10^{-180}$. *Hyb.* is end-to-end hybrid macro-F1 on the full split at a 20% budget and at the paper's budget (67.9%); encoder-only is 0.618$\pm$0.004. No rung meets the pre-registered winning criterion, and hybrid quality degrades monotonically with scale.

Verifier	Tot./Act.	F_1	Δ vs. enc.	Hyb. 20%	Hyb. paper
Qwen3.5-4B	4/4B	0.429	−0.181±0.010	0.600	0.479
Qwen3.5-9B	9/9B	0.396	−0.212±0.006	0.584	0.442
Qwen3.5-27B	27/27B	0.398	−0.192±0.009	0.582	0.452
Qwen3.5-122B-A10B	122/10B	0.399	−0.204±0.006	0.578	0.445
Qwen3.5-397B-A17B	397/17B	0.400	−0.212±0.006	0.571	0.444
Qwen3.7-Max	undisclosed	0.392	−0.221±0.008	0.566	0.439
Encoder only	–	0.618	–	0.618	0.618
Qwen3.5-4B LoRA	4/4B	0.608	−0.002±0.002	–	0.610

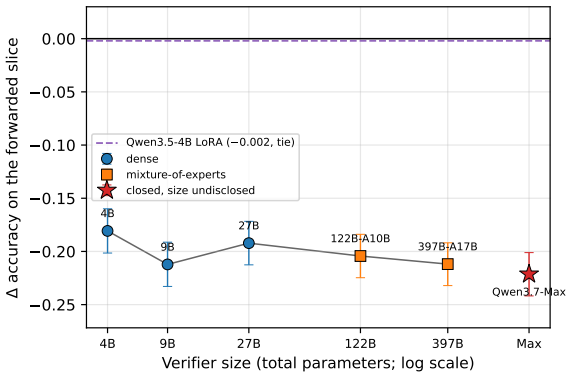


Figure 5: Slice Δ against verifier size, with 95% paired-bootstrap intervals. The gap neither closes nor trends; the dashed line is the supervised 4B verifier, the only configuration that reaches encoder parity.

that from the labelling convention, we ran the same ladder on the three other datasets in this study, holding prompt skeleton, decoding and parser fixed (Table 13, Fig. 6). Scale lifts all three. The contrast with CEDR is the sharpest: the verifiers reach 0.75-0.83 macro-F1 there while the transferred XLM-R encoder scores 0.577 on the same five classes, so these models read emotion well – in Russian, on a five-class Ekman subset – when the convention is one they can infer. (CEDR gold covers five of the seven Ekman classes, so all CEDR figures here average over those five; the encoder's seven-way average, which charges it for two structurally absent classes, would be 0.412.) Only GE, whose 28 emotions are folded into 7 Ekman classes by a fixed annotator convention, stays flat. Because the ladder rungs on AG News, ED and CEDR are standalone runs, these rows speak to verifier capability, not to gated escalation; the gated analysis remains GE (Table 12) and AG News (Appendix D.9).

Table 13: The same ladder, four tasks, standalone macro-F1. Scale lifts every task whose labels carry no dataset-specific grouping convention; GE is the exception, and the only task where the ladder declines.

Task	4B	9B	27B	122B	397B	Max	Δ
AG News (topic)	0.812	0.796	0.855	0.852	0.880	0.885	+0.073
CEDR (ru emotion)	0.746	0.752	0.784	0.805	0.809	0.827	+0.081
ED (emotion)	0.564	0.587	0.626	0.612	0.629	0.650	+0.086
GoEmotions	0.429	0.396	0.398	0.399	0.400	0.392	−0.037

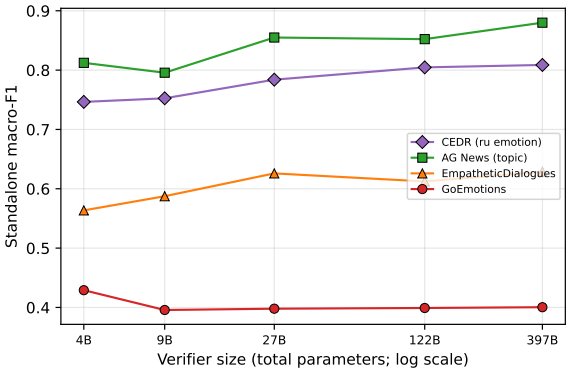


Figure 6: Standalone macro-F1 against verifier size on four tasks. GoEmotions is the only ladder that does not rise, which locates the difficulty in its label convention rather than in the models.

Why GoEmotions differs. Two measurements identify the cause. First, on GE the verifiers agree with *each other* far more than with the labels, and that gap is by far the largest of the four tasks (Table 14).

Table 14: Rung-to-rung agreement against rung-to-gold agreement, four tasks. *Model–model* is the mean pairwise agreement among the ladder rungs; *vs. gold* is their mean agreement with the label. A large gap means the models are consistent with each other and inconsistent with the annotation convention, which is what separates GoEmotions from every other task here.

Task	model–model	vs. gold	gap
GoEmotions	0.778	0.481	0.297
ED	0.860	0.653	0.207
AG News	0.929	0.850	0.079
CEDR	0.813	0.765	0.047

On GE the four largest rungs are unanimous on 67.7% of examples yet match the label only 56.0% of the time there, against 91.9% unanimity at 90.4% accuracy on AG News. The verifiers are consistent; they are calibrated to a different convention. Their errors fall along the seams of the official grouping: where the gold label is *surprise*, Qwen3.7-Max answers *neutral* 43% of the time (the mapping folds *realization*, *confusion* and *curiosity* into *surprise*), and where it is *anger* it answers *disgust* 21% of the time (the mapping folds *annoyance* and *disapproval* into *anger*).

Second, the macro-F1 decline along the GE ladder is not a broad degradation but one rare class. Decomposing 0.429 (4B) → 0.392 (Max) by class, *disgust* contributes −0.030 of the −0.038 total (79%): its per-class F1 falls 0.333 → 0.126 because the larger model predicts *disgust* on 11.0% of examples where the true share is 1.5%. Overall accuracy is flat (0.487 → 0.468) and the predicted label distribution actually moves *closer* to gold (total variation 0.199 → 0.139); the larger models simply draw an *anger/disgust* distinction the taxonomy collapses, and macro-averaging over a 1.5% class punishes that disproportionately. Treating the first gold label as the target is not the cause either: scoring leniently against any of an example's gold Ekman labels raises every rung by the same +0.023 to +0.027.

Can the convention be prompted? If the gap is a convention, telling the model the convention should close it. We re-ran a frozen random 2000-example subsample of the GE test split with the grouping made explicit in the prompt – *anger (also: annoyance, disapproval)*, *surprise (also: realization, confusion, curiosity)*, and so on – against the same subsample under the standard prompt. It helps, more for larger models, and nowhere near enough (Table 15).

Table 15: Spelling the GoEmotions grouping convention out in the prompt: standard prompt → taxonomy-explicit prompt, on a frozen 2000-example subsample of the GE test split. Every rung improves and the largest improves most, but none comes near the encoder's 0.586 accuracy on the forwarded slice.

Verifier	macro-F1	accuracy
Qwen3.5-4B	0.428 → 0.450	0.481 → 0.510
Qwen3.5-9B	0.380 → 0.396	0.476 → 0.500
Qwen3.5-397B-A17B	0.385 → 0.408	0.476 → 0.507
Qwen3.7-Max	0.367 → 0.406	0.450 → 0.502

The largest model gains the most (+0.039 macro-F1, +0.052 accuracy), consistent with it having drifted furthest along the *anger/disgust* seam, but even with the convention spelled out it reaches 0.406 against the encoder's 0.586 on the slice. A grouping convention can be described in a prompt only partially; the rest has to be learned from labelled data, which is exactly what the LoRA verifier does and why it is the one configuration that

ties.

Caveats. Four belong with these numbers. First, the 4B rungs and Gemma3-4B are local Ollama runs at Q4_K_M, whereas the API rungs are served at bf16/fp8; serving precision therefore varies along the ladder, though the direction of that effect would favour the larger rungs. Second, 9B and 27B are dense while 122B-A10B and 397B-A17B are sparse, so we report total and active parameters separately; the dense and sparse rungs nonetheless sit on the same flat line. Third, Qwen3.7-Max is closed-weights with an undisclosed size, which is why it is drawn detached from the scaling curve. Fourth, we could not evaluate the newer Qwen3.8-Max at all: its endpoint refuses the request with *"reasoning is mandatory for this endpoint and cannot be disabled"*. Forcing reasoning on would vary the decoding regime together with scale and multiply the per-call cost, which weighs against the cheap-escalation premise rather than for it. Because models behind an API can change silently, the exact model identifiers, the serving provider and the run date are recorded alongside every cached prediction

D.8 Escalation Beyond GoEmotions

The gated comparison of Appendix D.3 is run on GoEmotions. This appendix runs the identical protocol on all four tasks in the study, so that the mechanism is tested rather than assumed to transfer.

Protocol. One script produces every row (`cross_dataset_gates.py`). For each task and each encoder seed, we temperature-scale the encoder on that task's calibration split, rank the evaluation examples by margin, forward the lowest-margin B fraction, and compare the encoder against verifier accuracy on exactly those examples with the paired tests of Appendix D.12, combining seeds by Stouffer. The end-to-end hybrid answers forwarded examples with the verifier and everything else with the encoder, and is scored against the encoder-only macro-F1 of the same evaluation set. No new model calls were needed: the verifier predictions were already cached for all four tasks, and the encoder outputs that were missing (ED, CEDR) were produced by re-running the frozen checkpoints, each checked against its published full-coverage macro-F1 before use.

What differs between the tasks. GoEmotions uses the full 5427-example test split and three RoBERTa seeds; AG News the frozen 2000-example subsample and three seeds; ED the full 2542-example test split with the same three GoEmotions-trained encoders applied zero-shot; CEDR the 748 evaluation examples that remain after its 400-example calibration subsample, with the XLM-R encoder and the two calibration seeds used elsewhere in the paper. The verifier sets differ only in what was ever cached for that task.

Result. Table 16 gives the grid. On the three tasks with an in-domain encoder the verifier is worse on the forwarded slice, and the hybrid is below encoder-only, wherever the paired test has power. On ED, where the encoder is transferred, every zero-shot verifier is better on the forwarded slice at every budget and the hybrid beats encoder-only by a wide margin. The one verifier that does *not* help on ED is the LoRA verifier – supervised, but on GoEmotions' labelling convention rather than ED's – which is the same asymmetry that makes it the only verifier reaching parity on GoEmotions.

Three details are worth stating explicitly. First, AG News shows the effect *more* sharply than emotion does: an LLM scoring 0.812 standalone falls to 0.474 accuracy on the forwarded slice, and escalation costs 0.06 macro-F1 at a 20% budget, still 0.03 with the frontier verifier in the slot – so the selection effect is a property of encoder-first routing, not of affect annotation. Second, CEDR is the one task where the conclusion is budget-dependent, and the reason is that its encoder and its best verifier are nearly tied standalone (0.841 against 0.828): on 150 examples, a 20% budget forwards the paired test cannot separate them, and the loss only becomes significant from a 40% budget on. Its two seeds also share one encoder checkpoint, so its spread understates encoder uncertainty. Third, on ED the gain grows monotonically with verifier scale, from $\Delta = +0.142$ at 4B to $+0.252$ at Qwen3.7-Max, mirroring the standalone cross-task ladder of Appendix D.7: where the labelling convention is one the models can infer, scale helps behind the gate as well as in front of it.

D.9 A Non-Affective Closed-Label Task

Emotion is a favorable setting for the escalation hypothesis, which makes a negative result there informative; to separate mechanism from domain, we replicated the complete pipeline on AG News (Zhang et al., 2015), a 4-class news-topic dataset with no affective content, chosen because it is a closed-label task on which supervised encoders

Table 16: The gated comparison on all four tasks, one protocol. Rows indented under each task header are the size ladder, run behind that task's gate; the remaining rows are the verifier slots of the main search, which exist only where they were ever cached. *Standalone*: the verifier's own macro-F1 on that task's evaluation set, to be read against the encoder-only figure in the task header. Δ: verifier minus encoder accuracy on the forwarded slice, aggregated over seeds; † marks the ones a Stouffer-combined exact McNemar cannot separate from zero at $p < 0.05$, all others are significant. *Hyb.*: end-to-end macro-F1 of the escalating hybrid on the full evaluation set. CEDR is scored over the five Ekman classes its gold actually contains, and its two seeds share one encoder checkpoint – they differ only in which 400 examples calibrate it – so its spread understates encoder uncertainty and its 150-example slice at a 20% budget is the least powered cell in the table.

		budget 20%		budget 40%		budget 60%		budget 80%	
Verifier	Standalone	Δ	Hyb.	Δ	Hyb.	Δ	Hyb.	Δ	Hyb.
GoEmotions – encoder-only 0.618, $n = 5427$, 3 seeds									
Qwen3.5-4B zero-shot	0.429	−0.088	0.600	−0.128	0.557	−0.167	0.501	−0.198	0.450
Qwen3.5-9B	0.396	−0.128	0.584	−0.168	0.531	−0.200	0.467	−0.221	0.412
Qwen3.5-27B	0.398	−0.103	0.582	−0.141	0.532	−0.178	0.475	−0.209	0.421
Qwen3.5-122B-A10B	0.399	−0.121	0.578	−0.157	0.527	−0.192	0.467	−0.214	0.416
Qwen3.5-397B-A17B	0.400	−0.132	0.571	−0.165	0.521	−0.200	0.466	−0.221	0.417
Qwen3.7-Max	0.392	−0.138	0.566	−0.178	0.514	−0.211	0.460	−0.233	0.410
Qwen3.5-4B LoRA	0.608	+0.034	0.620	+0.011†	0.615	+0.002†	0.611	−0.005†	0.608
Voting k=3	0.416	−0.083	0.596	−0.122	0.553	−0.153	0.499	−0.189	0.440
Self-cons. k=5	0.415	−0.082	0.597	−0.126	0.552	−0.160	0.495	−0.192	0.438
Heterog. voting k=3	0.442	−0.081	0.599	−0.122	0.560	−0.153	0.507	−0.182	0.460
MultiAgentESC	0.341	−0.167	0.561	−0.216	0.497	−0.257	0.425	−0.284	0.363
AG News – encoder-only 0.936, $n = 2000$, 3 seeds									
Qwen3.5-4B	0.812	−0.277	0.879	−0.237	0.836	−0.179	0.824	−0.145	0.816
Qwen3.5-9B	0.796	−0.326	0.868	−0.286	0.812	−0.208	0.801	−0.160	0.798
Qwen3.5-27B	0.855	−0.207	0.894	−0.172	0.865	−0.124	0.859	−0.097	0.856
Qwen3.5-122B-A10B	0.852	−0.229	0.889	−0.178	0.862	−0.129	0.856	−0.101	0.853
Qwen3.5-397B-A17B	0.880	−0.158	0.904	−0.120	0.887	−0.087	0.883	−0.069	0.880
Qwen3.7-Max	0.885	−0.138	0.908	−0.106	0.893	−0.079	0.888	−0.063	0.885
CEDR, ru – encoder-only 0.841, $n = 748$, 2 seeds									
Qwen3.5-4B	0.749	−0.073†	0.846	−0.142	0.817	−0.147	0.795	−0.161	0.761
Qwen3.5-9B	0.749	−0.050†	0.843	−0.109	0.821	−0.117	0.800	−0.145	0.755
Qwen3.5-27B	0.776	−0.057†	0.848	−0.112	0.832	−0.116	0.814	−0.122	0.790
Qwen3.5-122B-A10B	0.806	−0.050†	0.849	−0.074	0.847	−0.080	0.835	−0.099	0.810
Qwen3.5-397B-A17B	0.811	+0.017†	0.859	−0.035†	0.849	−0.051	0.835	−0.063	0.819
Qwen3.7-Max	0.828	+0.003†	0.860	−0.047†	0.852	−0.047	0.848	−0.051	0.838
EmpatheticDial. – encoder-only 0.493, $n = 2542$, 3 seeds									
Qwen3.5-4B	0.564	+0.142	0.528	+0.116	0.556	+0.086	0.565	+0.061	0.563
Qwen3.5-9B	0.588	+0.183	0.536	+0.150	0.564	+0.124	0.581	+0.096	0.584
Qwen3.5-27B	0.626	+0.214	0.554	+0.182	0.597	+0.147	0.618	+0.113	0.625
Qwen3.5-122B-A10B	0.612	+0.201	0.549	+0.167	0.585	+0.137	0.605	+0.104	0.611
Qwen3.5-397B-A17B	0.629	+0.228	0.555	+0.193	0.596	+0.159	0.619	+0.125	0.628
Qwen3.7-Max	0.650	+0.252	0.559	+0.222	0.609	+0.184	0.636	+0.146	0.646
Qwen3.5-4B zero-shot	0.575	+0.133	0.534	+0.114	0.571	+0.080	0.579	+0.051	0.575
Qwen3.5-4B LoRA	0.534	+0.022†	0.516	−0.003†	0.527	−0.005†	0.532	−0.004†	0.533
Voting k=3	0.563	+0.144	0.533	+0.116	0.559	+0.082	0.561	+0.059	0.563
Self-cons. k=5	0.561	+0.150	0.534	+0.118	0.559	+0.084	0.561	+0.061	0.560
Heterog. voting k=3	0.550	+0.089	0.522	+0.071	0.546	+0.046	0.552	+0.028	0.549
MultiAgentESC	0.577	+0.171	0.533	+0.156	0.566	+0.127	0.576	+0.093	0.573

are very strong – the hardest possible case for our claim, since it leaves an LLM verifier a large error budget to recover. RoBERTa-base was fine-tuned with the paper's recipe on a stratified 20k train subsample (2 epochs, batch 32, lr 2×10^{-5}, seeds 42/43/44), temperature-scaled on a held-out 2k validation subsample, and zero-shot Qwen3.5-4B was run on a frozen stratified 2k test subsample (greedy decoding, no parse failures). Splits are frozen once and shared by all seeds, so only fine-tuning randomness varies and the LLM predictions are identical across seeds.

Table 17 sweeps the margin gate. Standalone, both systems are competent (encoder 0.936 ± 0.003, LLM 0.812 macro-F1), and the gap between them is smaller than the gap between their per-slice accuracies at every budget: escalation replaces encoder answers that are right 50-88% of the time with LLM answers that are right 44-68% of the time. End-to-end macro-F1 therefore falls monotonically with the escalation budget, from 0.936 with no escalation to 0.829 at a 50% budget. The effect is significant for budgets of 10% and above; at a 5% budget only 100 examples are forwarded, and the comparison is underpowered. The pattern is *sharper* here than on emotion – an LLM scoring 0.81 standalone falls to 0.44-0.68 accuracy on uncertainty-conditioned slices – so the selection effect is a property of encoder-first routing, not of affect.

Table 17: Margin-gate budget sweep on AG News (mean±std over 3 encoder seeds; $n = 2000$ test examples). Enc. and LLM are accuracies on the forwarded slice, Δ their difference, p the largest per-seed exact McNemar p-value, and F_1 the end-to-end macro-F1 of the escalating hybrid. Encoder-only macro-F1 is 0.936 ± 0.003.

Budget	n_{fwd}	Enc.	LLM	Δ	p	F_1
5%	100	0.500	0.437	−0.063±0.047	1.00	0.933±0.003
10%	200	0.622	0.473	−0.148±0.024	0.01	0.921±0.005
20%	400	0.751	0.474	−0.277±0.019	$<10^{-3}$	0.879±0.003
30%	600	0.818	0.545	−0.273±0.004	$<10^{-3}$	0.850±0.003
50%	1000	0.880	0.676	−0.204±0.002	$<10^{-3}$	0.829±0.002

D.10 Full Split-Conformal Results

Table 18 reports marginal split-conformal at three target miscoverage levels $\alpha \in \{0.05, 0.10, 0.20\}$ across GE, ED, and CEDR. GE and ED rows are mean±std over seeds $\{42, 43, 44\}$; CEDR rows are mean±std over seeds $\{42, 43\}$ (the seed-44 XLM-R checkpoint is left for future work). At larger α the threshold can yield empty sets, treated as abstentions; singleton acceptance is therefore not guaranteed to be monotone in α (explaining the slight CEDR drop from $\alpha = 0.10$ to $\alpha = 0.20$).

Table 18: Full split-conformal sweep. Cov.: empirical prediction-set coverage; Acc.: singleton-set rate; $F_{1,\text{acc}}$: macro-F1 on accepted examples. GE/ED: 3 seeds; CEDR: 2 seeds.

Dataset	α	cov.	acc.	$F_{1,\text{acc}}$
GE	0.05	0.947±0.011	0.242±0.042	0.807±0.022
GE	0.10	0.905±0.008	0.366±0.023	0.817±0.005
GE	0.20	0.813±0.009	0.655±0.019	0.705±0.012
ED	0.05	0.967±0.003	0.010±0.003	0.132±0.006
ED	0.10	0.916±0.003	0.071±0.009	0.256±0.017
ED	0.20	0.823±0.001	0.251±0.001	0.525±0.010
CEDR	0.05	0.960±0.002	0.662±0.015	0.660±0.007
CEDR	0.10	0.902±0.024	0.894±0.047	0.619±0.020
CEDR	0.20	0.794±0.010	0.883±0.007	0.618±0.006

D.11 Confidence-Score Ablation

Within the encoder-first selective router, we compare three confidence scores – margin, maximum probability, and an entropy-derived confidence – on GoEmotions at the best-found threshold ($\theta \approx 0.86$), together with the two threshold extremes (Table 19). Full-test macro-F1 and cost are identical across all of them by construction: the router only abstains, so no choice of score changes the encoder's own predictions, and every variant issues the same one encoder call per example. What the score changes is which examples are accepted, and there margin separates correct from incorrect encoder predictions best. We therefore use margin throughout.

Table 19: Confidence-score and threshold ablation on GoEmotions (seed 42). Full-test macro-F1 is 0.623, and cost is one encoder call for every row (abstention never overrides the encoder), so only the accepted set moves. *Cov@5%*: coverage at the 5% accepted-error budget; *Fwd.*: fraction of examples the gate forwards.

Variant	$F_{1,\text{sel}}$	Cov@5%	Fwd.
Margin (default)	**0.826**	0.223	66.7%
Maximum probability	0.815	0.216	57.6%
Entropy-derived	0.667	0.217	73.5%
Low threshold θ=0.30	0.689	0.223	20.8%
High threshold θ=0.99	0.248	0.223	91.0%

D.12 Rejected-Slice Statistical Testing

For each encoder seed, we form the seed-specific rejected slice under the validation-selected margin threshold and compare paired correctness indicators between the encoder and each verifier

using exact McNemar's test. We convert each exact p-value to a signed z-score using the sign of the verifier-minus-encoder accuracy difference, then combine the seed-level scores with Stouffer's method. Table 20 gives the per-seed inputs behind every combined p we report, so the aggregation can be checked rather than taken on trust. Since all non-LoRA verifier comparisons remain significant at $p < 10^{-3}$, the rejected-slice conclusion is unchanged under Bonferroni correction.

Table 20: Per-seed exact McNemar on the rejected slice (GoEmotions), for the verifier slots of the main search; the ladder rungs and Gemma3-4B are tested identically, and their per-seed rows are released with the ladder artifact. n: size of that seed's slice; b/c: discordant pairs where only the encoder / only the verifier is correct; concordant pairs carry no information and drop out. Δ is verifier minus encoder accuracy. LoRA-Qwen is the one verifier whose per-seed tests all fail to reject, which is what the combined $p = 0.68$ of Table 7 aggregates.

Verifier	Seed	n	Δ	b	c	p
Qwen3.5-4B LoRA	42	3657	−0.002	490	482	0.82
	43	3461	+0.000	443	444	1.00
	44	3936	−0.004	479	463	0.63
combined (Stouffer)						0.68
Qwen3.5-4B zero-shot	42	3657	−0.177	1135	488	$<10^{-3}$
	43	3461	−0.173	1068	470	$<10^{-3}$
	44	3936	−0.192	1247	490	$<10^{-3}$
Voting k=3	42	3657	−0.165	1108	506	$<10^{-3}$
	43	3461	−0.160	1035	483	$<10^{-3}$
	44	3936	−0.180	1215	505	$<10^{-3}$
Self-cons. k=5	42	3657	−0.169	1102	483	$<10^{-3}$
	43	3461	−0.166	1036	462	$<10^{-3}$
	44	3936	−0.186	1221	489	$<10^{-3}$
Heterog. voting k=3	42	3657	−0.161	1073	483	$<10^{-3}$
	43	3461	−0.158	1007	462	$<10^{-3}$
	44	3936	−0.175	1182	495	$<10^{-3}$
MultiAgentESC	42	3657	−0.270	1353	366	$<10^{-3}$
	43	3461	−0.264	1282	369	$<10^{-3}$
	44	3936	−0.282	1496	385	$<10^{-3}$

D.13 Preliminary AutoMAS Observations (Scoping Only)

We adapted AutoAgents (Chen et al., 2024a) and EvoAgent (Yuan et al., 2025) to Qwen3.5-4B over subsets of GoEmotions and ED. Each example invokes a multi-step reasoning workflow with long-generation prompts, not short classification calls. End-to-end wall-clock latency is $\approx$ 46 s (AutoAgents, $\approx$ 5 calls) and $\approx$ 138 s (EvoAgent, $\approx$ 9 calls) per example. On available subsets they underperform simpler baselines: AutoAgents reaches $F_1 = 0.376$ on GE ($n = 1450$) and 0.527 on ED; EvoAgent reaches 0.239 and 0.409 respectively ($n = 450$). We list these as scoping evidence, not primary baselines, because their call budget and latency place them outside the low-overhead regime targeted by the encoder-first selective router. These numbers are not used in the main Pareto-front claims; they only motivate excluding full AutoMAS workflows from the low-overhead verifier search.

E Use of AI assistants

The authors used Claude (Anthropic) and DeepSeek to assist with typesetting, phrasing of selected passages, and structural editing of this paper. The original text of this paper was spell- and grammar-checked and slightly smoothed out using Grammarly. All experimental design, analysis, results, and conclusions are the authors' own.