JOURNAL OF INFORMATION SCIENCE AND ENGINEERING XX, XXXX-XXXX (2024)
DOI: XX.XXX.XXXX.XX

AspectRoute: Cost-Aware Confidence-Gated Escalation to LLMs for Aspect-Based Sentiment Analysis of Entertainment Reviews

M. N. VIJAI KUMAR, P. VINOTHIYALAKSHMI AND G. VISHWESH+
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, Tamil Nadu, India
E-mail: Mnvijaikumar7@gmail.com; vlakshmi@svce.ac.in; visweshsubramanian@outlook.com

Sentiment analysis of user-generated entertainment reviews, the textual feedback that drives discovery on streaming platforms, has predominantly relied on document-level classification, obscuring the fine-grained opinions users hold about distinct content aspects such as narrative quality, acting, visual aesthetics, and pacing. Large language models (LLMs) classify such aspects well but are conventionally applied uniformly to every input. We propose AspectRoute, which instead decides *per aspect mention* whether LLM inference is warranted: aspect terms are extracted by a keyword pipeline with contextual windowing, and each aspect-context pair is escalated to an LLM only when a RoBERTa encoder is insufficiently confident, with all remaining pairs classified locally at no API cost. On 41,315 aspect mentions from 9,916 IMDB movie reviews, escalating 57.5% of pairs attains a Coverage-F1 of 0.823 (review-level bootstrap 95% CI $[0.818, 0.828]$) against 0.826 $[0.820, 0.831]$ for escalating every pair, at 58% of the cost and with higher coverage (74.4% vs. 67.0%); the same 57.5% escalated at random yields 0.774, indicating that the gain derives from *which* pairs are escalated rather than from LLM use alone. We note that this escalation rate exceeds our 35–40% design target, so the realised saving is smaller than the design would in principle allow; we report the full cost–quality curve across thresholds. On the SemEval-2014 Task 4 test split, where labels are genuinely aspect-level, AspectRoute reaches macro-F1 0.772 (restaurants) and 0.698 (laptops), significantly above the zero-shot encoder on both domains ($p = 1.1 \times 10^{-6}$; $p = 0.004$) without task-specific training. We introduce Inter-Aspect Sentiment Divergence (IASD) and find that 32.3% of multi-aspect reviews express contradictory sentiments across aspects, with positive-Performance / negative-Narrative the dominant pattern. We characterise the contribution as an empirically characterised selective-inference policy, together with a measurement of what escalation costs and what it buys, rather than as a new classification algorithm. Evaluation against binary document-level IMDB labels remains an inherent mismatch with three-class aspect-level prediction, which the SemEval evaluation is included to address; a purpose-built streaming-platform benchmark remains a priority for future work.

Keywords: aspect-based sentiment analysis, large language models, selective inference, confidence-gated escalation, entertainment reviews, explainability, cost-efficient NLP, inter-aspect sentiment divergence

1. INTRODUCTION

Streaming platforms host vast repositories of textual feedback encoding nuanced opinions about content quality. Unlike traditional product reviews, entertainment reviews are characterised by multidimensional evaluations: a single review may praise a show's cinematography while criticising its pacing. Conventional sentiment analysis assigns one

Received (date); revised (date); accepted (date).
+Corresponding author.

label to the whole review, obscuring this. “The visuals were breathtaking but the plot was a complete mess” receives a neutral aggregate sentiment, concealing both assessments. Lexicon-based tools such as VADER [1] are fast but struggle with sarcasm, negation, and domain-specific vocabulary. Transformer-based models like BERT [2] and RoBERTa [3] offer superior contextual understanding but at significant computational cost. GPT-family LLMs demonstrate impressive few-shot sentiment classification but are too expensive to apply uniformly at scale [4].

We identify three underaddressed gaps. First, while aspect-based sentiment analysis (ABSA) is well-studied for product and restaurant reviews [5, 6], the entertainment domain has a distinct aspect taxonomy (narrative arc, rewatchability, cultural authenticity) not covered by existing benchmarks. Second, LLM-based pipelines are conventionally applied *uniformly*: every input receives the same expensive treatment regardless of whether it needs it, and the question of which inputs actually require LLM inference is rarely posed, let alone measured. Third, explainability is rarely integrated into sentiment pipelines [7].

Contributions. We propose AspectRoute:

1. A *selective-inference policy* for aspect-level sentiment analysis that decides, per aspect mention, whether LLM inference is warranted, using encoder confidence as a proxy for linguistic difficulty.
2. An empirical characterisation of what that decision costs and what it buys: escalating 57.5% of mentions recovers Coverage-F1 0.823 against 0.826 for escalating all of them, at 58% of the cost, while escalating the same proportion at random yields only 0.774.
3. An entertainment-review aspect taxonomy and keyword-based extraction pipeline covering seven aspect categories, with per-path token-level attribution.
4. The IASD metric quantifying contradictory sentiment across aspects within a single review, which we find in 32.3% of multi-aspect reviews.

We do not claim that selective routing between models is itself new: RouteLLM [8], CARGO [9] and related work [10] established it for general-purpose query routing, and LLMs [4] and explainability [7] have each been applied to ABSA. Our claim is narrower and, we believe, defensible: that applying such a policy at the *aspect level within a single document*, and measuring empirically what the escalated model does and does not contribute, has not to our knowledge been reported. The framework is modular in the classifier used for non-escalated mentions, and we report two instantiations, one lexical and one encoder-based, to show that the escalation policy, rather than any particular local classifier, carries the result. Streaming-platform validation is left for future work; IMDB serves here as an entertainment-review proxy.

2. LITERATURE SURVEY

Work relevant to this paper spans lexical sentiment analysis, transformer encoders, LLM-based classification and adaptive routing; we take them in turn and close by identifying the gap that motivates selective inference at the aspect level. Beginning with lexical methods, VADER [1] remains widely cited for its speed and interpretability on social media text. Its limits are well documented: Dhanalakshmi et al. [11] compared VADER with logistic regression for structured review classification, though neither addressed aspect-level granularity, and Hossen and Dev [12] obtained only marginal gains from domain-specific lexicon extensions, constrained by the difficulty rule-based approaches face with context-dependent language. More broadly, Drus and Khalid [13] catalogued over 200 sentiment studies and found that lexicon-based methods consistently underperform machine learning on linguistically complex data. This motivates our use of a lexical classifier

only where an encoder indicates the input is straightforward.

Turning to neural encoders, RoBERTa [3] improved upon BERT's [2] pre-training strategy. Applied to ABSA, Ansar et al. [5] demonstrated BERT-based aspect extraction achieving F1 > 0.85 on restaurant datasets, though not on entertainment content, while Paneru et al. [14] combined RoBERTa embeddings with CNNs at the document level. Encoders are strong and inexpensive once trained, but their confidence estimates, which we exploit as a routing signal rather than only as a classification output, are seldom used to decide where further computation is warranted.

With regard to LLM-based approaches, Simmering and Huoviala [4] showed that fine-tuned GPT-3.5 achieves state-of-the-art F1 on SemEval-2014 Task 4 ABSA at substantially higher cost than encoder models, framing precisely the cost-accuracy tension that Kumar et al. [15] identify as a central open challenge. Relevant to our routing signal, Herrera-Poyatos et al. [16] documented how sarcastic and ironic expressions cause inconsistent LLM classifications across inference runs. A representative recent system is that of Gaeta et al. [17], who combine role-playing, few-shot and chain-of-thought prompting for propaganda detection in text; as in most such pipelines, prompt design is the central methodological lever and the LLM is applied uniformly to every input. Our concern is the complementary question of which inputs require LLM inference at all.

Within ABSA specifically, Mai and Le [6] showed that modelling aspect interactions improves classification, Chen [18] reviewed multimodal ABSA, and Rodriguez-Ibanez et al. [19] emphasised the necessity of domain-specific adaptation. Closest to our work, Perikos et al. [7] combined ABSA with SHAP and LIME explanations on transformer models; like the LLM pipelines above, they apply a single model uniformly and do not target entertainment content.

On adaptive routing, RouteLLM [8] showed that routing between weak and strong models using learned preference data can achieve large cost reductions while maintaining GPT-4-level performance on selected benchmarks, CARGO [9] extended this with embedding-based regressors for multi-model routing, and Zhang et al. [10] proposed semantic entropy for routing uncertainty estimation. These frameworks route whole queries in general-purpose settings. Applying the same principle at the granularity of individual aspect mentions within a single document, where dozens of routing decisions may be made per review, is to our knowledge unexplored.

Finally, Abelard and Sibaroni [20] applied LDA for multi-aspect sentiment analysis on Netflix app reviews using SVM, without transformers, LLMs, routing or explainability. It remains the closest prior study on multi-aspect sentiment for streaming content, and we extend it with neural models, selective escalation and aspect-level attribution.

Taken together these strands leave a gap: lexical methods are cheap but brittle, encoders are cheap and strong but their uncertainty is discarded, LLMs are accurate but applied indiscriminately, and routing work operates on whole queries rather than individual aspects. What is missing is a policy deciding per aspect mention whether LLM inference is warranted, with measurement of what it buys. We claim novelty for the policy and its characterisation, not for any component.

3. PROPOSED METHODOLOGY

3.1 System Overview

AspectRoute operates as a four-stage pipeline: (1) *Aspect Extraction*: raw review $\rightarrow$ aspect-opinion pairs via keyword matching with contextual windowing; (2) *Confidence-Gated Routing*: each pair routed to VADER or LLM based on RoBERTa encoder confidence (Eq. 1); (3) *Sentiment Classification*: augmented VADER or LLM produces a three-class label; (4) *Explainability*: feature-level SHAP for VADER-routed predictions

and perturbation-based token attribution for a sampled subset of LLM-routed predictions.

3.2 Dataset

We evaluate on the IMDB dataset [21] (50,000 binary-labelled movie reviews), sampling 10,000 for tractable LLM evaluation; 9,916 of these yield at least one aspect mention. We acknowledge that IMDB provides document-level binary labels, creating an evaluation mismatch with three-class aspect-level prediction. Coverage-F1 (Section 4.2) is introduced specifically to address this; binary F1 is reported as a secondary reference metric. We define an entertainment-review aspect taxonomy with seven categories: Narrative, Performance, Visual, Audio, Emotional Impact, Cultural Relevance, and Technical. These were arrived at by the authors through iterative reading of approximately 500 reviews. We note explicitly that this was not a formal coding procedure: no codebook was fixed in advance, no independent coders were used, and no inter-annotator agreement was computed. The taxonomy should therefore be read as a working scheme rather than a validated instrument. An executable annotation protocol for validating it, specifying corpus, schema, procedure and a Fleiss' κ target, is provided with the released code.

3.3 Keyword-Based Aspect Extraction

We implement aspect extraction using a curated lexicon of 200+ terms mapped to seven categories, applying contextual windowing to extract the sentence or clause containing each matched term. At most one mention per category per review is retained, and the first matching term per category in document order is used. This approach prioritises precision and interpretability. A neural variant using RoBERTa-base with BIO sequence labelling is described as a future direction (Section 8). All experiments reported in this paper use keyword-based extraction.

3.4 Confidence-Gated Routing

Let c_i denote the maximum softmax confidence from the RoBERTa sentiment classifier for aspect pair i. Routing is:

$$\text{route}(i) = \begin{cases} \text{VADER} & c_i \geq \tau \wedge \neg \text{sarcasm}(i) \\ \text{LLM} & \text{otherwise} \end{cases} \tag{1}$$

with $\tau = 0.85$ selected by a grid search over $[0.70, 0.95]$ at 0.05 intervals on the IMDB sample (Coverage-F1 as the selection criterion); no SemEval labels were used in threshold selection. We note as a limitation that this grid search was conducted on the same IMDB sample used for evaluation rather than a held-out split; the threshold is a single coarse hyperparameter, but a fully held-out tuning protocol is left to future work, and the SemEval results (Section 5.5), which use this fixed τ without any domain-specific retuning, provide an out-of-domain check. A secondary trigger escalates any mention whose context matches one of 20 hand-curated surface patterns intended to flag sarcasm. We describe this accurately as a surface-pattern heuristic rather than as sarcasm detection: matching is case-insensitive substring containment with no contextual disambiguation, so it cannot distinguish a sarcastic use of a phrase from a sincere one. Section 5.8 reports what this component actually contributes, and Section 7 treats its weaknesses as a limitation.

We use the RoBERTa encoder as a *linguistic-ease proxy*: a high confidence indicates that the aspect context is linguistically straightforward, with unambiguous polarity and little sarcasm or negation, and that escalation is unlikely to change the outcome, whereas a low confidence flags the genuinely hard cases that justify the cost of an LLM call.

The framework is agnostic to which classifier handles non-escalated mentions, and we report two instantiations: the augmented VADER pipeline of Section 3.5, and the encoder's own prediction, which is already available at no extra cost since the confidence score must be computed for routing regardless. The submitted version adopted the

lexical variant on interpretability grounds; Section 5 shows the encoder variant is substantially more accurate at identical cost and remains interpretable through token-level attribution. We therefore report it as the recommended configuration. The escalation policy and threshold are identical across both.

3.5 Augmented VADER Pipeline

For high-confidence routing, VADER is enhanced with: (1) a domain lexicon of 847 entertainment-review terms with manually assigned sentiment scores (*bingeworthy*: $+3.2$, *plot hole*: -2.8, *formulaic*: -1.9); (2) aspect-aware context windowing; and (3) grid-search-optimised thresholds (negative: $c \leq -0.15$; positive: $c \geq +0.20$; neutral: otherwise).

3.6 LLM Classification via Structured Prompting

Each LLM call provides the aspect-containing context window and the aspect category label, and instructs the model to return exactly one of three sentiment labels while explicitly attending to sarcasm, implicit sentiment, and negation. The classification is *zero-shot*: no in-context examples are supplied, and the prompt is identical across aspect categories (the full template is given in the supplementary material and ships with the released code). Temperature is set to 0.1. GPT-4o-mini (`gpt-4o-mini-2024-07-18`) serves as the backbone at \$0.15/1M input and \$0.60/1M output tokens. Each response is requested as a JSON object containing the sentiment label, a self-reported confidence, and a one-sentence rationale; malformed responses fall back to neutral with confidence 0.5. A local DistilBERT classifier (`lxyuan/distilbert-base-multilingual-cased-sentiments-student`) is provided as a fallback for environments without API access; all results reported in this paper use the GPT-4o-mini backbone.

3.7 Per-Path Explainability

For VADER-routed predictions, SHAP values [22] are computed over the augmented lexicon features. For LLM-routed predictions, perturbation-based attribution masks input tokens and measures changes in the returned confidence as an approximate attribution signal; this is an input-level approximation, as the API does not expose the LLM's internal weights. SHAP/attribution is computed on a sampled subset of predictions; where this is infeasible, averaged attention weights from the RoBERTa router serve as a coarse fallback. We do not claim faithful internal attribution for every prediction, and we do not claim SHAP on the LLM path.

4. EXPERIMENTAL SETUP AND BASELINES

All experiments use Python 3.11, PyTorch 2.1, HuggingFace Transformers 4.36 and SHAP 0.43, run on a MacBook Air M1 (8 GB RAM) with random seed 42 throughout. The escalated path uses GPT-4o-mini (snapshot `gpt-4o-mini-2024-07-18`) at temperature 0.1 in all reported runs.

4.1 Baselines and Their Implementation

We compare against four reference points, each described with the implementation details that determine its cost and behaviour.

VADER-Only applies standard VADER to the full review text, with no aspect extraction; it is free and serves as the lower bound. **VADER-Aspect** applies the augmented pipeline of Section 3.5 to aspect context windows, isolating the contribution of windowing and the domain lexicon; also free.

Twitter-RoBERTa applies `cardiffnlp/twitter-roberta-base-sentiment-latest` [23], part of the TimeLMs line [24], zero-shot to every pair, with no escalation. It is the key reference point, being the same encoder that produces our routing

signal: comparing against it isolates what escalation adds over using the encoder's predictions everywhere. It runs locally, free.

LLM-all escalates all 41,315 pairs using the same prompt, snapshot and temperature as the escalated path. It is the upper reference for quality and worst case for cost, replacing the proxy used in the submitted version.

RoBERTa-FT is `roberta-base` fine-tuned on the in-domain SemEval-2014 training split, used in the SemEval evaluation only. It represents the supervised alternative that our fine-tuning-free setting is measured against.

4.2 Evaluation Metrics

We report binary macro-F1, precision, recall, and accuracy (neutral $\rightarrow$ negative). We introduce Coverage-F1 to evaluate the quality of non-neutral predictions separately from coverage:

$$\text{Coverage-F1} = \text{macro-F1}\{(y_i, \hat{y}_i) : \hat{y}_i \neq \text{neutral}\} \tag{2}$$

with coverage $= |\{i : \hat{y}_i \neq \text{neutral}\}|/N$. Coverage-F1 rewards accurate committed predictions and lets a system abstain on genuinely ambiguous cases without being penalised as if it had answered wrongly. Because it rewards abstention by construction it must be read together with the coverage figure, which we report in every table. Binary F1 is a secondary reference.

Significance is assessed via bootstrap (2,000 resamples) for confidence intervals and continuity-corrected McNemar tests for paired comparison. For IMDB, where multiple mentions originate from the same review, resampling is at the *review level*, drawing whole reviews with all their mentions, so within-review correlation is preserved and intervals are not artificially narrowed. For SemEval, where most sentences contribute a single instance, resampling is at the instance level; where a sentence contributes several this is an approximation and intervals may be marginally optimistic. All McNemar tests use only items where both systems commit, making them coverage-matched by construction.

We also report Inter-Aspect Sentiment Divergence (IASD):

$$\text{IASD} = \frac{|\{r : \exists a_1, a_2 \in A_r,\ \{\hat{y}_{a_1}, \hat{y}_{a_2}\} = \{+, -\}\}|}{|\{r : |A_r| \geq 2\}|} \tag{3}$$

Here A_r is the set of aspect mentions extracted from review r, and $\hat{y}_a$ the predicted label for mention a. The numerator counts reviews containing at least one mention predicted positive and at least one predicted negative; the denominator counts reviews with at least two extracted mentions. In words, IASD is the proportion of multi-aspect reviews whose aspects carry conflicting sentiment. A review praising the acting while criticising the plot contributes to the numerator; one whose aspects are all positive, or all neutral, does not.

4.3 Reproducibility and a Correction to the Original Run

During preparation of this revision we identified a fault in the original pipeline: the LLM classification function caught every exception and returned a neutral label with confidence 0.5, so calls that failed under API rate limiting were silently recorded as abstentions rather than as errors. Re-running the same model with the same prompt over the same items raises the commit rate from 6.7% to 62.2%, with 99.7% agreement wherever both runs committed ($n = 1{,}460$), confirming the difference is failed calls rather than model variability. All 23,763 escalated predictions were regenerated with corrected error handling (zero failures, zero missing) and every IMDB result recomputed. SemEval was unaffected: it required far fewer calls and its escalated items committed at 80–88%.

Code, lexicons, cached and regenerated predictions, and a harness that reproduces every figure reported here from cache without API calls are available at `https://github.com/vijaikumar-7/AspectRoute`.

5. RESULTS AND DISCUSSION

5.1 Overall Classification Performance

Table 1 reports results on 41,315 aspect mentions from 9,916 IMDB reviews. All figures derive from the regenerated predictions described in Section 4.3.

The central comparison is between escalating selectively and escalating everything. Escalating the 57.5% of mentions on which the encoder is least confident attains a Coverage-F1 of 0.823 $[0.818, 0.828]$, against 0.826 $[0.820, 0.831]$ for escalating all 41,315 mentions, with overlapping intervals, at 58% of the cost, and with higher coverage (74.4% vs. 67.0%). A paired McNemar test on the 27,377 mentions where both commit still favours full escalation ($b = 135$, $c = 257$, $p < 10^{-9}$), so we describe this as near-parity rather than equality: selective escalation gives up a small but real amount of accuracy in exchange for a substantial cost reduction.

Against the zero-shot encoder applied everywhere, the same model that produces the routing signal at no cost, selective escalation improves Coverage-F1 from 0.771 to 0.823 at essentially unchanged coverage (74.6% vs. 74.4%), with a decisive paired test ($p \approx 10^{-66}$). The LLM is consulted on a minority of mentions and improves the outcome on them, while the encoder handles the rest for free.

Coverage-F1 rewards abstention by construction, so a system that commits less often can score higher without being better. We therefore report coverage in every row, since the configurations compared here sit within eight points of one another, and restrict all paired tests to mentions where both systems commit, which removes the asymmetry entirely.

Under binary mapping, VADER-Only scores highest among the cheap baselines because applying VADER to full review text approximates document-level classification and so matches IMDB's document-level ground truth. This is an artefact of the label mismatch, and precisely the limitation the SemEval evaluation addresses.

Table 1. Overall results on IMDB ($n = 41{,}315$ aspect mentions from 9,916 reviews, $\tau = 0.85$). Coverage-F1 is macro-F1 over committed (non-neutral) predictions; intervals are 95% review-level bootstrap. Cost is measured from observed token counts at GPT-4o-mini pricing. AspectRoute rows differ only in the classifier used for non-escalated mentions.

Configuration	Cov-F1	95% CI	Cov. %	Commit acc.	Cost/1K
VADER-Only	0.652	[.645, .658]	79.8	65.6	$0.000
Twitter-RoBERTa (zero-shot)	0.771	[.764, .777]	74.6	77.1	$0.000
AspectRoute (lexical local path)	0.786	[.780, .792]	72.2	78.6	$0.123
AspectRoute (encoder local path)	**0.823**	**[.818, .828]**	**74.4**	**82.4**	**$0.123**
LLM-all (escalate everything)	0.826	[.820, .831]	67.0	82.8	$0.213

5.2 Inter-Aspect Sentiment Divergence

Of 9,486 multi-aspect reviews, 3,066 contain at least one positively and one negatively labelled aspect, giving IASD = 32.3%. The dominant pattern is positive Performance with negative Narrative (587 cases, 19.1% of divergent reviews), followed by positive Visual with negative Narrative (580, 18.9%) and positive Performance with negative Emotional Impact (525, 17.1%). Narrative appears as the negative pole in three of the five most frequent patterns, consistent with the observation that viewers frequently praise execution while criticising the story.

That roughly a third of multi-aspect reviews carry internally conflicting sentiment substantiates the premise motivating this work, and remains conservative since extraction retains at most one mention per category per review.

5.3 Escalation Rate and Cost

At $\tau = 0.85$, 42.5% of mentions are handled locally and 57.5% escalated, against a design target of 35–40%. We report this gap plainly: the realised saving is smaller than the design would in principle allow. The cause is domain mismatch: the encoder was pre-trained on Twitter text and is systematically less confident on movie-review prose (mean confidence 0.780, median 0.819).

Costs are computed from observed token counts (mean 132.3 input and 52.2 output tokens per call) at GPT-4o-mini pricing. Escalating everything costs \$0.213 per 1,000 reviews; selective escalation costs \$0.123, a 42.5% reduction. The figures in the submitted version derived from an assumed per-call constant rather than measured usage and were too high by roughly an order of magnitude; the relative reduction, being a routing ratio, is unchanged.

These absolute figures scale linearly with backbone price: a few cents per thousand reviews with an inexpensive model, but substantial with a frontier model priced one to two orders of magnitude higher. They are also API costs only; the local path additionally avoids per-call network latency, which dominated wall-clock time in our runs.

5.4 What Does Escalation Actually Contribute?

Of the 23,763 escalated mentions the LLM commits on 62.2% and abstains on the rest. On those it commits to, accuracy against IMDB document labels is 78.5%, whereas the encoder achieves 85.9% on the non-escalated mentions it handles. The escalated set is harder by construction, so this gap reflects the difficulty of the routed items rather than a deficiency of the LLM.

The LLM's predictions are also substantially independent of the encoder's: on escalated mentions the two disagree 44.4% of the time. Escalation therefore supplies a genuinely different inductive bias on exactly the mentions where the encoder is least reliable, which is what the ablation in Section 6 confirms, since the same volume of escalation applied at random recovers far less of the benefit.

The submitted version reported a 93.3% abstention rate on escalated items and characterised calibrated abstention as the LLM's primary contribution. That figure was an artefact of the fault described in Section 4.3; with correct predictions the rate is 37.8% and the claim does not hold. We withdraw it.

5.5 SemEval-2014 Task 4: Aspect-Level Gold Standard

The IMDB evaluation is limited by document-level ground truth. We therefore evaluate on the SemEval-2014 Task 4 test split [25], where labels are genuinely aspect-level: 994 aspect instances for restaurants and 543 for laptops (not sentences; the official test split contains 800 sentences per domain). Following a common three-class simplification [4, 7] we map `conflict` to `neutral`. The threshold τ is unchanged and no SemEval labels were used for tuning of any kind.

Table 2 reports the results. On restaurants, selective escalation reaches macro-F1 0.772, above both the zero-shot encoder (0.686) and the in-domain fine-tuned encoder (0.702), with accuracy of 0.919. On laptops it reaches 0.698, again above both references. Paired McNemar tests confirm significance against the zero-shot encoder on both domains ($p = 1.1 \times 10^{-6}$ restaurants; $p = 0.004$ laptops) and against the lexical local-path variant ($p < 10^{-6}$ both domains).

The comparison against RoBERTa-FT deserves care: point estimates favour selective escalation on both domains, but intervals overlap and we did not retain per-item predictions from the fine-tuning run, so no paired test is possible. We therefore claim only that a fine-tuning-free policy is *competitive with* in-domain supervised training. This is meaningful nonetheless, since RoBERTa-FT consumed 2,338 labelled examples on restaurants

and AspectRoute consumed none.

Table 2. SemEval-2014 Task 4, held-out test split. Three-class macro-F1 with 95% bootstrap CIs. RoBERTa-FT is *roberta-base* fine-tuned on the in-domain training split; all other systems use no task-specific training.

Method	Restaurants F1 [95% CI]	Laptops F1 [95% CI]
VADER-Aspect	0.574 [.538, .610]	0.553 [.510, .595]
Twitter-RoBERTa (zero-shot)	0.686 [.647, .725]	0.684 [.636, .728]
RoBERTa-FT (in-domain, supervised)	0.702 [.662, .741]	0.653 [.606, .698]
AspectRoute (lexical local path)	0.703 [.662, .739]	0.658 [.615, .701]
AspectRoute (encoder local path)	**0.772** [.727, .810]	**0.698** [.649, .745]

Two observations replicate across domains: the neutral class is hardest for every method, confirming minority-class detection as the binding difficulty; and VADER-Only and VADER-Aspect coincide because SemEval sentences are already short and aspect-specific, so context windowing adds nothing here.

5.6 Worked Attribution Examples

Attribution is computed as described in Section 3.7. We give two examples, one successful and one not, to show both how the signal is read and where the pipeline fails.

In a correctly classified Emotional Impact mention (IMDB label negative, system output negative), attribution concentrates on the evaluative terms: *horrible* ($+0.277$), *stupid* ($+0.204$), *badly* ($+0.060$) and *silly* ($+0.032$), with function words contributing negligibly. The explanation is legible and matches the intuitive reading of the sentence.

The second is a failure, included deliberately. For the Visual aspect in "we lovers of 'Lost' can see the lack of respect the producers of this **stunning** series have with the fans", attribution concentrates on *Unfortunately* ($+0.276$) and *lack* ($+0.229$) while the matched keyword *stunning* contributes *negatively* (-0.075). The attribution is behaving correctly; extraction has failed, since the sentence criticises the producers rather than the visuals. This substantiates the extraction limitation in Section 7 and shows attribution to be useful for diagnosis, not only presentation.

5.7 Sensitivity to the Escalation Threshold

Table 3 reports Coverage-F1, coverage and cost across the threshold range. The values are exact rather than interpolated: the set of mentions escalated at any $\tau \leq 0.85$ is a subset of those escalated at $\tau = 0.85$, so every required prediction already exists. We do not report $\tau = 0.90$ or 0.95, at which 17% and 33% respectively of the escalated mentions lack a regenerated prediction.

Coverage-F1 rises monotonically and smoothly with τ, from 0.804 to 0.823, while cost rises from \$0.072 to \$0.123 per 1,000 reviews. The result therefore does not depend on a narrow optimum, an important reassurance given that τ was selected on the same IMDB sample used for evaluation (Section 3.4), and τ is best understood as an operating point on a cost–quality curve: at $\tau = 0.70$ the system still exceeds the zero-shot encoder (0.804 vs. 0.771) while escalating a third of mentions at a third of the cost.

5.8 What Triggers Escalation?

Escalation has two triggers, low encoder confidence and a sarcasm pattern match, and it is worth reporting how much each contributes, since the second is a crude heuristic. Of the 23,763 escalated mentions, 22,225 (93.5%) were escalated on low confidence alone, 993 (4.2%) matched both criteria, and 545 (2.3%) were escalated *only* because of a pattern match. The sarcasm heuristic therefore independently accounts for 1.3% of all aspect

Table 3. Sensitivity to the escalation threshold τ, recommended configuration. Exact values, computed from regenerated predictions.

τ	Cov-F1	Coverage %	% escalated	Cost/1K
0.70	0.804	76.3	33.6	\$0.072
0.75	0.811	76.0	40.6	\$0.087
0.80	0.817	75.4	48.4	\$0.103
0.85	0.823	74.4	57.5	\$0.123

mentions, and the system's routing behaviour and cost profile are driven overwhelmingly by encoder confidence.

The patterns themselves are less flattering. Five of the twenty (*oh great*, *wow just wow*, *thanks for nothing*, *never seen that before*, *how creative*) never matched once across 41,315 contexts, and 98.1% of matches came from eight generic discourse markers (*of course*, *certainly*, *as if*, *obviously*, *apparently*, *clearly*, *shocking*, *supposedly*) ordinarily used sincerely. The component functions as a hedge-word detector, not a sarcasm detector. We note that the patterns cited illustratively in the submitted version were unrepresentative of its behaviour, and treat replacement by a learned classifier as future work. A second LLM backbone on the same 4,000 escalated mentions changed committed accuracy by two points (77.3% vs. 79.4%), with over 95% agreement on jointly-committed labels.

6. ABLATION STUDY

Table 4 isolates the contribution of each element, reusing the regenerated predictions throughout.

A2 is the most informative: escalating the same 57.5% of mentions *at random*, at identical cost, yields 0.774 against 0.823. Almost half the benefit comes not from using an LLM but from using it on the right mentions, which is this paper's central claim stated as an ablation. A4 removes escalation entirely (0.771), close to random routing. A4b substitutes the lexical classifier for the encoder and drops to 0.652, so the local path matters once escalation is fixed. A5 is the full-escalation reference, now measured; the proxy used in the submitted version estimated 0.852, so it was optimistic.

A1 (no domain lexicon) and A3 (no aspect windowing) are omitted: both require re-running extraction rather than re-using predictions, and we prefer to omit them than report figures inherited from the faulty run.

Table 4. Ablations ($n = 41{,}315$). Δ is the change in Coverage-F1 relative to the recommended configuration. A2 holds cost constant and varies only which mentions are escalated.

Configuration	Cov-F1	Δ
AspectRoute (recommended)	0.823	—
A2: random escalation at the same 57.5% rate	0.774	-0.049
A4: no escalation (encoder only)	0.771	-0.053
A4b: no escalation (lexical only)	0.652	-0.172
A5: full escalation (measured)	0.826	$+0.002$
Lexical local path variant	0.786	-0.037

7. LIMITATIONS

The most consequential limitation concerns the original evaluation rather than the method. As described in Section 4.3, a silent error-handling fault recorded failed API calls as abstentions, and every IMDB result has been regenerated. The regenerated run

completed without failures and agrees with the original wherever the original succeeded, but the episode illustrates how readily an unmonitored failure mode can masquerade as a substantive finding.

The evaluation mismatch on IMDB remains: binary document-level ground truth cannot fully validate three-class aspect-level predictions, and Coverage-F1 rewards abstention by construction. We mitigate this by reporting coverage alongside every score, restricting paired tests to jointly-committed items, and including the SemEval evaluation (Section 5.5). Neither substitutes for a purpose-built entertainment benchmark with aspect-level annotations and inter-annotator agreement; we provide an executable protocol for constructing one with the released code, and regard it as the highest-priority next step.

The taxonomy and all three hand-built resources, namely the 200+ term extraction lexicon, the 847-term domain lexicon and the 20 sarcasm patterns, were built by the authors without a formal coding protocol, independent coders or inter-annotator agreement, and we cannot supply that retrospectively. The sarcasm heuristic is demonstrably poor (Section 5.8): it cannot distinguish sarcastic from sincere usage, five of twenty patterns never fire, and its precision and recall are unmeasurable without sarcasm ground truth. Its small influence on routing bounds but does not excuse this.

The escalation threshold was selected on the same IMDB sample used for evaluation rather than a held-out split; Section 5.7 mitigates this by showing performance varies smoothly rather than peaking at the chosen value, and the SemEval results use the same fixed τ without retuning.

Further limitations: English only; keyword extraction misses implicitly referenced aspects and selects the first matching term rather than the most salient, as Section 5.6 illustrates; the escalated path depends on external API availability and pricing; attribution is computed on a sampled subset; and the comparison against in-domain supervised training rests on overlapping intervals without a paired test.

8. CONCLUSION

We presented AspectRoute, a selective-inference policy for aspect-based sentiment analysis that decides, per aspect mention, whether LLM inference is warranted. On 41,315 IMDB aspect mentions, escalating the 57.5% on which a RoBERTa encoder is least confident attains Coverage-F1 0.823 against 0.826 for escalating every mention, at 58% of the cost and with higher coverage; escalating the same proportion at random yields 0.774, so the benefit derives from *which* mentions are escalated. On the SemEval-2014 test split, where labels are genuinely aspect-level, the policy reaches macro-F1 0.772 (restaurants) and 0.698 (laptops), significantly above the zero-shot encoder on both and competitive with in-domain supervised fine-tuning, using no task-specific training. IASD shows 32.3% of multi-aspect reviews carry internally conflicting sentiment.

We characterise the contribution as a selective-inference policy together with a measurement of what escalation costs and what it buys, not as a new classification algorithm. Several findings are unfavourable: full escalation remains marginally more accurate, the sarcasm heuristic does not detect sarcasm, and absolute savings are small at current prices for inexpensive models. We report them because a policy for deciding when to spend on inference is only useful if the accounting is honest.

Future directions: a purpose-built entertainment benchmark with aspect-level annotations, for which the released code specifies a protocol; a fine-tuned sequence labeller to replace keyword extraction, which the failure case in Section 5.6 motivates directly; a learned sarcasm classifier; domain-adapted encoder fine-tuning to reduce the escalation rate toward the design target; selective-risk curves; and extension to open-weight back-

bones whose pricing would alter the cost analysis.

REFERENCES

1. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th International AAAI Conf. Web and Social Media (ICWSM)*, Ann Arbor, MI, USA, 2014, pp. 216–225.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
3. Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
4. P. F. Simmering and P. Huoviala, "Large language models for aspect-based sentiment analysis," *arXiv preprint arXiv:2310.18025*, 2023.
5. W. Ansar, S. Goswami, A. Chakrabarti, and B. Chakraborty, "An efficient methodology for aspect-based sentiment analysis using BERT through refined aspect extraction," *Journal of Intelligent & Fuzzy Systems*, Vol. 40, no. 5, 2021, pp. 9627–9644.
6. L. Mai and B. Le, "Joint sentence and aspect-level sentiment analysis of product comments," *Annals of Operations Research*, Vol. 300, no. 2, 2021, pp. 493–513.
7. I. Perikos *et al.*, "Explainable aspect-based sentiment analysis using transformer models," *Big Data and Cognitive Computing*, Vol. 8, no. 11, 2024, p. 141.
8. W. Ong *et al.*, "RouteLLM: Learning to route LLMs with preference data," *arXiv preprint arXiv:2406.18665*, 2024.
9. A. Barrak, Y. Fourati, M. Olchawa, E. Ksontini, and K. Zoghlami, "CARGO: A framework for confidence-aware routing of large language models," *arXiv preprint arXiv:2509.14899*, 2025.
10. T. Zhang, A. Mehradfar, S. Avestimehr, and D. Dimitriadis, "Leveraging uncertainty estimation for efficient LLM routing," *arXiv preprint arXiv:2502.11021*, 2025.
11. P. Dhanalakshmi, G. Ashish Kumar, B. Sai Satwik, K. Sreeranga, A. Tharun Sai, and G. Jashwanth, "Sentiment analysis using VADER and logistic regression techniques," in *Proc. 2023 Int. Conf. Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, Coimbatore, India, 2023, pp. 139–144.
12. M. S. Hossen and N. R. Dev, "An improved lexicon based model for efficient sentiment analysis on movie review data," *Wireless Personal Communications*, Vol. 120, 2021, pp. 535–544.
13. Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Computer Science*, Vol. 161, 2019, pp. 707–714.
14. B. Paneru, B. Thapa, and B. Paneru, "Sentiment analysis of movie reviews: A flask application using CNN with RoBERTa embeddings," *Systems and Soft Computing*, Vol. 7, 2025, p. 200192.
15. M. Kumar, L. Khan, and H.-T. Chang, "Evolving techniques in sentiment analysis: A comprehensive review," *PeerJ Computer Science*, Vol. 11, 2025, p. e2592.
16. A. Herrera-Poyatos *et al.*, "An overview of model uncertainty and variability in LLM-based sentiment analysis: Challenges, mitigation strategies, and the role of explainability," *Frontiers in Artificial Intelligence*, Vol. 8, 2025, p. 1609097.
17. A. Gaeta, V. Loia, A. Lorusso, F. Orciuoli, and A. Pascuzzo, "Towards a LLM-based intelligent system for detecting propaganda within textual content," *Computers and Electrical Engineering*, Vol. 128, 2025, p. 110765.
18. T. Chen, "A review of multimodal aspect-based sentiment analysis," *Advances in Engineering Innovation*, Vol. 16, no. 6, 2025, pp. 43–51.
19. M. Rodríguez-Ibáñez, A. Casañez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems with Applications*, Vol. 223, 2023, p. 119862.
20. A. R. Abelard and Y. Sibaroni, "Multi-aspect sentiment analysis on netflix application using latent dirichlet allocation and support vector machine methods," *INFOTEL*, Vol. 13, no. 3, 2021, pp. 128–133.
21. A. L. Maas *et al.*, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting ACL*, Portland, OR, USA, 2011, pp. 142–150.
22. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vol. 30, Long Beach, CA, USA, 2017, pp. 4765–4774.
23. J. Camacho-Collados *et al.*, "TweetNLP: Cutting-edge natural language processing for social media," in *Proc. EMNLP 2022 (Syst. Demos)*, Abu Dhabi, UAE, 2022, pp. 38–49.
24. D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-Collados, "TimeLMs: Diachronic language models from Twitter," in *Proc. 60th Annu. Meeting ACL (Syst. Demos)*, Dublin, Ireland, 2022, pp. 251–260.
25. M. Pontiki *et al.*, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 27–35.