

HIGH-LOAD BUDGETED CATEGORIZATION OF CUSTOMER CARE CALLS: AN ENCODER-LLM CASCADE SOLUTION

Anonymous authors

Paper under double-blind review

ABSTRACT

Industry operators route tens of millions of customer-call summaries a year into 100+ fine-grained, long-tailed categories, and at that volume the choice of model per call is itself a cost decision: cheap fine-tuned encoders are unreliable on ambiguous calls, while a strong LLM resolves them but costs one to two orders of magnitude more. We present a deployed hybrid encoder-LLM cascade. A confidence gate on a fine-tuned encoder answers the calls it can and escalates only the rest to the LLM, handing it just the handful of labels the encoder could not separate rather than the full taxonomy. We formulate the two coupled decisions—whether to escalate, and how much of the label space to expose—as one calibrated policy, of which the deployed fixed-gate, fixed-shortlist system is the special case we measure.

The economics are the central result. Because the encoder clears roughly 87% of calls on its own, the cascade cuts total cost by more than 90% against running the LLM over the full label list on every call, and shortlisting the escalated call labels cuts their tokens again—the margin that makes the pipeline viable at this volume rather than prohibitive.

Shortlisting also carries a prompt-design lesson: the optimized shortlisted prompt is not the full-list prompt with fewer options but a different prompt, and one written for the full taxonomy transfers poorly when reused unchanged—scoring below the full list on human-gold routed calls even when the true label is present. Tuned for the shortlisted regime, grounding each candidate in a synthesized definition lifts conditional accuracy from 42.7% to 63.5%. Past the prompt, the last lever is shortlist recall—set by ranker quality, not by shortlist sizing or further prompt tuning.

1 INTRODUCTION

Production classification systems in customer operations must assign every inbound call a single fine-grained intent label drawn from a large, long-tailed taxonomy. This setting is common to any high-volume call center or customer-interaction channel, across industries. In the deployment we study, an industry operator categorizes tens of millions of call-transcript summaries per year into 100+ fine-grained categories whose frequencies span several orders of magnitude. Two facts make this hard at once. First, the label space itself is a source of error: the taxonomy contains many semantically close categories whose boundaries are subtle, and even human annotators disagree on them, so the “gold” labels used for evaluation are themselves contested on a non-trivial fraction of calls. Second, the cost/accuracy tension is acute: a fine-tuned encoder classifies a call for a fraction of a cent but is unreliable on ambiguous inputs, whereas a strong instruction-tuned LLM resolves ambiguity well but is one to two orders of magnitude more expensive per call—far too costly to run on every call at this volume.

The dominant response in the literature is a **cascade**: let the cheap encoder answer the easy calls and escalate only the hard ones to the LLM. A parallel line of work makes LLM classification over large label spaces feasible by **shortlisting**—a cheap upstream model proposes a handful of candidate labels and the LLM chooses among only those. Crucially, prior work tunes these two

controls *separately*: cascade research decides *whether* to escalate while holding the LLM’s task fixed, and shortlist research decides *which* labels to expose while invoking the LLM on every input. Neither treats the two decisions as one coupled, budgeted object.

Our starting observation is simple: **sometimes the encoder is sure of the label, and sometimes it is confused among a few.** Conformal prediction makes this operational. A single calibration step turns the encoder’s probabilities into a per-call *confusion set*—the labels that remain plausible once we account for how often the encoder’s confidence is wrong. When one label survives, the encoder is “sure” and its answer stands; the LLM is never invoked. When several survive, the call is escalated and the confusion set *becomes* the LLM’s candidate shortlist. A single coverage level α thus fixes the escalation rate, the shortlist length, and the cost at once, while keeping the true label on the shortlist at a controlled rate—enforced *per class-frequency band* so rare categories are not silently squeezed out. We call this the **Conformal Shortlist Cascade (CSC)**, and note that classic fixed-threshold, fixed- k routing (with k the shortlist size) is a special case of this policy class.

Deploying a cascade of this shape also surfaces a prompt-design lesson that shaped the paper. **The shortlisted prompt is a different prompt from the full-list one, not just a shorter version of it: a picker written for the full taxonomy transfers poorly when reused unchanged on the shortlist, and can score below the full list even when the true label is present.** On the routed hard slice, a full-list-style prompt applied unchanged to the shortlist scores below what the same picker reaches on the full list, even though the true label is present the large majority of the time. The effect is not a recall artifact—we condition on the true label being present—and survives our checks for label contestability and candidate ordering; it is a property of the *prompt*, and a shortlist-adapted prompt reverses it (§7). Two levers therefore govern the routed slice: adapting the picker prompt, the large one; and **shortlist recall**, the complementary ceiling that CSC’s per-band guarantee is designed to control.

Together these observations shape the paper. We formulate encoder-to-LLM classification as budgeted joint control of two coupled per-call decisions—whether to invoke the LLM, and how much of the label space to expose to it—for which confidence-threshold routing with a fixed shortlist is the fixed- (τ, k) special case. Our evaluation instantiates the framework at exactly this production operating point (a fixed margin threshold τ and shortlist size k); the fixed-gate numbers we report are that special-case instance, not a demonstration of adaptive-conformal superiority. Empirically, over 100+ long-tailed categories, we (i) show that the shortlisted prompt must be tuned as a prompt in its own right—a full-list-style prompt transfers poorly to the shortlist, a conditional-accuracy gap on human gold that survives our label-contestability and ordering checks; (ii) isolate definition-grounding as the strongest prompt lever for recovering conditional selection on the routed slice; and (iii) show that shortlist recall—set by ranker quality, not the sizing rule—imposes a separate hard ceiling on end-to-end accuracy, with per-call dynamic sizing showing algorithm-dependent, not-yet-settled gains on the routed branch (a selection-effect hypothesis, argued from our data, may explain why consistent gains are hard to realize there). To our knowledge this is the first systematic study of jointly budgeted escalation and adaptive candidate-set construction in a deployed encoder→LLM cascade for single-label classification over 100+ long-tailed classes, with measured relative token and cost economics.

2 RELATED WORK

Cascades and routing (the “whether to escalate” axis). Confidence-gated escalation from a cheap model to an expensive one is well established: FrugalGPT (Chen et al., 2023) and Tabi (Wang et al., 2023) route by calibrated confidence. Decision-theoretic treatments characterize the optimal deferral rule and its limits: Jitkrittum et al. (2023) show plain confidence thresholds are suboptimal under specialist experts or label noise, and Bouchard (2026) give a decision-theoretic characterization of the cost–quality frontier that does not treat candidate-set restriction as a decision variable. Budgeted threshold selection (Kotte, 2026; Valkanas et al., 2025) picks escalation thresholds under cost constraints; learned deferral (Kondadadi & Ortega, 2026) trains the gate from uncertainty signals. All of these keep the LLM’s task fixed—the label space it sees is not a control variable.

Candidate-space restriction (the “what to expose” axis). A second line makes LLM classification over large spaces feasible by shortlisting. The closest single mechanism is the Dual-Expert

paradigm (Cheng et al., 2024), in which a fine-tuned expert proposes a fixed top- k set (with no accept/escalate gate) and an LLM selects among them; retrieval-based variants (Dev et al., 2025; Zhang et al., 2025; Zhu & Zamani, 2023; Vandemoortele et al., 2025) retrieve candidates from a large label space and let the LLM pick. CROQ (Vishwakarma et al., 2025) prunes multiple-choice options to a conformal set and shows LLMs are more accurate on fewer options; related work (Lu et al., 2024) documents that many/ambiguous options degrade LLM decisions. These establish the *set-size* $\leftrightarrow$ *accuracy* trade-off we build on, but they invoke the LLM for **every** input—no cheap-model termination, no cost budget, no long-tail structure.

Long-tailed calibration and conformal sets. Conformal prediction for long-tailed classification (Ding et al., 2025) and macro-coverage guarantees (Bhattacharyya et al., 2026) produce prediction sets that are valid *per class*, trading set size against rare-class coverage; calibration under many classes (Le Coz et al., 2024) and post-hoc imbalance recalibration (Tian et al., 2020) support class-aware gating. This machinery gives us tail-valid shortlists—but it has never been used to *jointly* drive escalation and construct an LLM candidate space.

Position. No verified prior work occupies the intersection of (i) a selective gate that terminates easy calls, (ii) adaptive restriction of the label space exposed to the LLM, (iii) a supervised encoder base, (iv) explicit long-tail / per-class analysis, and (v) a measured cost budget at production scale. Dou et al. (2026) have the conformal *gate* but discard the set; CROQ passes conformal sets to the LLM but never gates and has no cost model. CSC is precisely their union, plus a per-band recall guarantee, studied in a deployed system.

3 PROBLEM SETUP AND EVALUATION PROTOCOL

Task. Single-label classification of a customer-call summary into one of 100+ fine-grained categories. Category frequencies span several orders of magnitude, from high-volume intents (e.g., a common payment-related intent) to rare tail classes.

Pipeline. A fine-tuned encoder (DeBERTa-large) produces a softmax over the 100+ categories and a confidence signal (the top-two margin $\gamma = p_1 - p_2$, the gap between the two largest softmax masses); a decision rule either accepts the encoder’s top-1 or escalates the call, exposing a shortlist of candidate labels to a strong instruction-tuned LLM (the *picker*) which selects the final label (Figure 1). The decision rule is left pluggable on purpose—the paper studies it as a design axis (conformal at level $1 - \alpha$, fixed (τ, k) , dynamic- k)—so the figure names it as a single node rather than hard-coding one method. Conformal is the primary instantiation and carries the per-frequency-band recall guarantee of at least $1 - \alpha$.

Data. The encoder is trained on a silver-labeled set ($\sim 38k$ train, $\sim 12.7k$ validation, $\sim 12.7k$ test; stratified 60/20/20) whose labels come from an LLM labeling process—formed by majority consensus across several frontier LLMs under one shared prompt template; we treat the silver set **only** as encoder training data and make no claim about its validity here. Evaluation uses a **human-gold set** of ~ 900 calls that was held out of encoder training and is therefore truth-bearing. Silver-labeled test numbers are reported only as consistency checks (the encoder was trained on silver, so silver agreement is optimistic).

Metric. Our primary picker metric is **accuracy conditioned on the true label being in the shortlist** (accuracy@ k given true-in-top- k). Conditioning on presence isolates the picker from shortlist recall, so any measured drop is a picker effect rather than a recall artifact. We also report recall@ k , end-to-end path accuracy (recall@ $k \times$ picker-accuracy@ k), macro-F1, and per-band coverage. Throughout, lowercase k denotes the shortlist size and K the number of categories. Sample sizes referenced below relate as follows: the human-gold set is ~ 900 calls (~ 860 true-in-top-5), of which the low-confidence *hard slice* is $n \approx 96$ ($\pm \sim 10pp$) and carries the headline human-scored conditional-accuracy numbers; the larger routed low-confidence slice ($\sim 1,030$ calls, scored against silver consensus where human labels are unavailable) is used only for the error diagnostics of §6.

Leakage-free authoring. Any prompt content that could be tuned to data (synthesized category definitions, pairwise exclusion rules; §7) is derived **only** from the validation split—validation predic-

tions come from a held-out HPO checkpoint—then frozen and applied unchanged to test, exactly as a promoted production prompt would behave. All operating points (α ; margin threshold τ ; shortlist size k) are selected on validation and reported once on test.

4 METHOD: THE CONFORMAL SHORTLIST CASCADE (CSC)

Confusion sets from calibration. Given the encoder’s probabilities $\hat{p}(y | x)$ over the 100+ labels (with x the call summary and y a candidate label), split-conformal calibration produces, for a chosen miscoverage level α , a prediction set $C_\alpha(x)$ that contains the true label with probability at least $1 - \alpha$. We read $C_\alpha(x)$ as the encoder’s *confusion set*. If $|C_\alpha(x)| = 1$ the encoder is “sure”: accept its label and never call the LLM. If $|C_\alpha(x)| > 1$ the encoder is “confused”: escalate the call and pass exactly $C_\alpha(x)$ as the LLM’s candidate shortlist.

Budgeted formulation and subsumption. For each call we choose to *accept* the encoder label or *escalate with a k -candidate set*, to maximize expected accuracy subject to an inference-cost budget. The probability a call is answered correctly given escalation with k candidates factorizes as

$$P(\text{correct} \mid \text{escalate}, k) = R@k(x) \cdot A(x, k), \quad (1)$$

shortlist recall $R@k$ (rising in k) times picker accuracy given the shortlist A (which, as §5 shows, can *fall* in k). The Lagrangian of the budgeted problem yields a value-of-escalation rule; **fixed- (τ, k) routing is the special case** in which the same τ and k are applied to every call. This subsumption is what makes fixed-routing baselines meaningful rather than straw men. In the deployed system we evaluate, the reported operating point *is* this fixed- (τ, k) special case (a top-5 shortlist at a tuned margin gate); conformal calibration is the framework it sits within, and realized per-band coverage is reported as validation of the construction rather than as a separately tuned system (see §11).

One knob. Given the two hand-tuned thresholds of fixed routing—a confidence threshold τ on the margin γ and a shortlist size k —conformal calibration collapses them into a single dial: the coverage level α becomes the system’s only knob. Lower α (more caution) makes more calls “confused” and lengthens confusion sets $\rightarrow$ more escalation, and more tokens; higher α is cheaper. Sweeping α traces the entire accuracy–cost frontier.

Tail-valid guarantee. Standard conformal in long-tailed settings forces a bad choice between small sets with poor rare-class coverage and per-class-valid but huge sets. Using prevalence-adjusted / macro-coverage scores, CSC enforces $\geq 1 - \alpha$ coverage **within each class-frequency band**, so the true label survives the shortlist at the target rate even for rare categories—shortlist recall becomes a design parameter rather than an empirical hope, tail included. This is a property of the construction under exchangeability; realized per-band coverage on test is reported as design validation, not as a separate empirical claim.

Operational safeguards. To bound worst-case cost and keep behavior safe we (i) cap the confusion-set size at $k_{\max}$ (which can exclude the true label and void the $\geq 1 - \alpha$ guarantee—an explicit guarantee-for-cost trade); (ii) constrain or validate the picker’s output to lie inside the shortlist, falling back to the encoder’s top-1 on violation; and (iii) estimate the $\alpha \rightarrow$ expected-cost mapping on validation and report its extrapolation error. The decision rule is deliberately pluggable (Figure 1): conformal at level $1 - \alpha$ is the primary instantiation and the one that carries the per-band guarantee, while fixed- (τ, k) and dynamic- k are alternative points in the same policy class that we compare against.

5 THE SHORTLIST PENALTY

The motivation for shortlisting is cost: presenting 5 candidates instead of all 100+ shortens the picker’s prompt sharply—in our deployment the shortlisted prompt uses $\approx 0.14\times$ the input tokens of the full-list prompt ($\approx 2.9\times$ cheaper per escalated call; §9), which is the whole reason a budgeted system shortlists at all. The implicit assumption is that this saving is accuracy-neutral whenever the true label is on the shortlist. It is not.

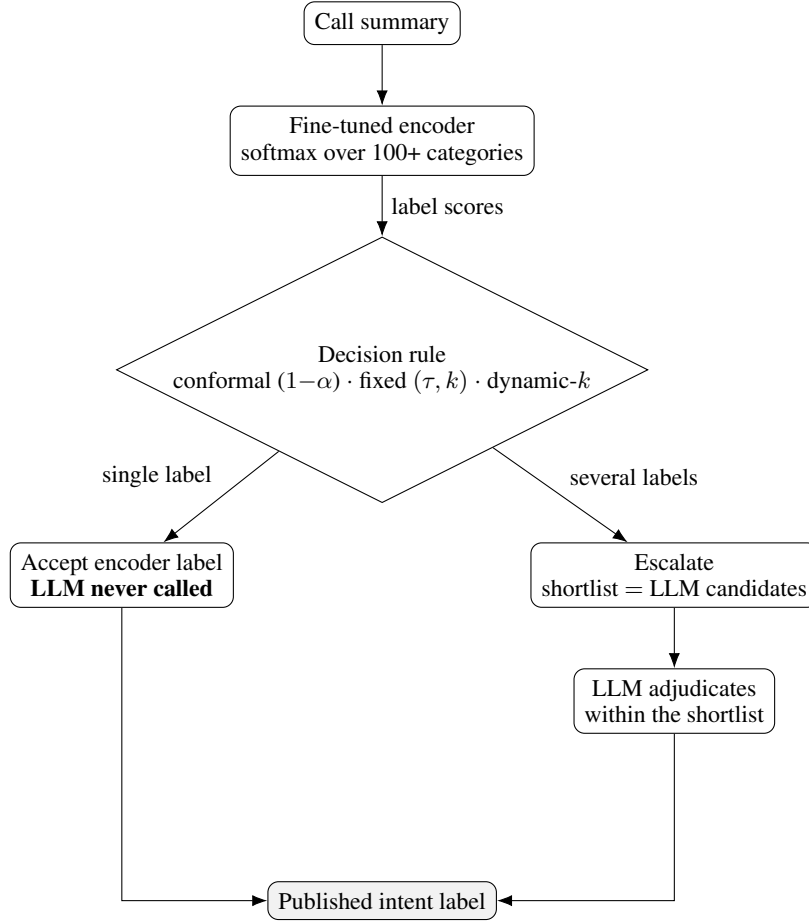


Figure 1: **The hybrid encoder-LLM cascade.** A fine-tuned encoder scores each call summary; its label scores pass through a decision rule that yields one of two outcomes. A single surviving label is accepted directly and the LLM is never invoked; several surviving labels escalate the case and become the LLM’s candidate shortlist. The decision rule is a design axis (conformal at level $1 - \alpha$, fixed (τ, k) , or dynamic- k); under conformal calibration a single coverage level sets the escalation rate, shortlist length, and cost at once, with a per-frequency-band recall guarantee of at least $1 - \alpha$.

On the human-gold routed hard slice ($n \approx 96$, $\pm \sim 10$ pp; picker at temperature 0), we compare the **same** picker on the **same** calls with and without the shortlist, scored only where the true label is present so that any gap is a picker effect rather than a recall artifact. A picker prompt written for the full 100+-label taxonomy, applied unchanged to the 5-candidate shortlist, scores only 42.7%—below the 52.1% the same picker reaches on the full list (Table 1), even though the true label is on the shortlist over 90% of the time. The raw gap is directional at this slice size, but it is one-sided and, as §7 shows, does not reverse into an advantage until the prompt is redesigned for the shortlisted regime.

We do not claim to have invented the phenomenon that fewer/ambiguous options can affect LLM decisions (CROQ and Lu et al. (2024) document related effects); our contribution is to **quantify it inside a cost-gated cascade over a 100+-class long-tailed taxonomy, under a leakage-free conditional metric on human-gold labels**, to show it is not explained by label contestability or candidate ordering (§6), and to show it is a property of the prompt that a shortlist-adapted design reverses (§7).

Table 1: Five-arm prompt study scored against **human** gold labels on the routed low-confidence slice ($n \approx 96$; accuracy@5 given true-in-top-5). Each arm is an incremental, leakage-free change layered on the previous (validation-derived, frozen, applied to test). Definition-grounding (Arm D) produces the strongest conditional accuracy among the tested shortlist prompts. Percentages are directional ($\pm \sim 10$ pp at this sample size).

Arm	Change	Accuracy
<i>Full category-list prompt, all 100+ (no shortlist)</i>		52.1%
A	Baseline — candidate names only	42.7%
B	+ calibrated framing + justified “Other”	51.0%
C	+ contrastive reasoning + forced best-to-worst ranking	56.2%
D	+ synthesized per-category CORE definitions	63.5%
E	+ targeted pairwise exclusion rules	58.3%

6 THE GAP IS A PROMPT EFFECT, NOT AN ARTIFACT

The shortlist transfer gap is robust to our checks for label contestability and candidate ordering, and is not fully explained by either. These diagnostics run on the larger routed low-confidence slice ($\sim 1,030$ calls), where silver-consensus labels stand in where human gold is unavailable; we read them as corroborating diagnostics, not as headline estimates.

Not label noise: even on the slice’s unanimously-labeled (clean) calls—those the labeling panel agreed on—the shortlist picker is still wrong on nearly a third (68% correct), so the gap cannot be an artifact of contestable ground truth. To bound how much label noise *could* explain, we apply a deliberately strict criterion: a routed call counts as genuinely ambiguous only when the full-list picker, shown all 100+ categories, independently lands on the *same* non-gold label as the shortlist picker. If an unrestricted picker makes the identical “mistake,” the shortlist did not cause it and the gold is the likelier culprit; by that conservative floor only 5.5% of routed calls qualify. *Not ordering:* conditional accuracy is flat across ranks 1–3 (60.0/59.2/60.2%) with a drop only at rank 4 (52.2%), and an order-agnostic full-context model makes the same confusions. The residual errors concentrate on a handful of specific near-twin category pairs rather than spreading uniformly; having ruled out label noise and ordering, we read the gap as a property of the unadapted prompt, which §7 addresses directly.

7 RECOVERING THE GAP: DEFINITION-GROUNDED SHORTLISTS

Can prompt design recover the penalty? The picker prompt must be *redesigned for the shortlisted regime*: a prompt written for the full taxonomy transfers poorly once the model must instead discriminate among a handful of near-neighbor candidates. In a leakage-free five-arm study (validation-authored, frozen, applied to test; arm-by-arm human-gold results in Table 1), the strongest single lever is grounding each shortlisted candidate in a synthesized 1–2 sentence “typical call” definition, built offline from ~ 30 validation calls per category.

On the paired human-gold routed slice ($n \approx 96$), adapting the picker prompt from a full-list-style baseline to the shortlist-specific, definition-grounded prompt raises conditional accuracy from 42.7% to 63.5%—a +20.8-point observed improvement on the same routed calls. Prompt adaptation is therefore a primary lever for conditional selection, not a marginal one, and Arm D is the strongest tested configuration on this human-gold routed slice. On the broader golden set ($n=860$, dominated by high-confidence calls) all five arms cluster at 74–82% and the shortlisted arms are competitive with the full-context reference, indicating that the poor baseline transfer is concentrated on the low-confidence routed slice rather than uniformly observed. What prompt adaptation *cannot* do is recover a true label the ranker never placed on the shortlist, which motivates shortlist recall as the complementary lever.

Table 2: Recall@ k on the held-out human-gold set: a supervised encoder ranker vs. RRF retrieval. To reach 90% recall, retrieval needs $k = 15$ but the encoder needs $k = 4$; for 95%, $k = 31$ vs. $k = 8$. Δ is the encoder’s recall advantage over retrieval at each k .

k	RRF (retrieval)	Encoder (DeBERTa)	Δ
1	.423	.732	+.308
5	.753	.931	+.177
10	.857	.962	+.105
15	.907	.972	+.065

8 RECALL IS THE COMPLEMENTARY ACCURACY CEILING

Once the picker prompt is fixed, end-to-end path accuracy is $R@k \times A@k$ —and the picker **cannot** recover a label absent from the shortlist. So the dominant remaining *controllable* lever is shortlist recall. Its *primary* driver is ordering quality, while the rule that sets the shortlist size k is only a *secondary* one—and that sizing rule takes two forms: a *static* k fixed across all routed calls, or a *dynamic* per-call k read off the encoder’s confidence. Concretely, in the end-to-end decomposition on human gold, the best-arm result is $0.931 \times 0.819 \approx 76\%$ on the full set and $0.80 \times 0.635 \approx 51\%$ on the hard slice; the gap between the 63.5% shortlist-adapted conditional-picker accuracy and the 51% end-to-end figure is attributable to shortlist recall and cannot be closed by picker-prompt work alone.

Ordering quality dominates. A supervised, correctness-aware ranker crushes retrieval orderings. Ranking all 100+ categories by the encoder’s per-label probability vs. a strong hybrid retrieval baseline (RRF over BM25 + dense) on the held-out human-gold set is shown in Table 2.

The recall ceiling was a **weak-ordering problem, not a cutoff problem**—a supervised ranker shifts the whole curve left. Residual misses are concentrated (50% of misses come from 8 categories, 80% from 19) and fall into three archetypes: irreducible catch-alls (*Other*, *Unclear inquiry*), confusable near-synonyms (the reranker/hard-negative target, $\sim 65\%$ of misses), and high-volume low-miss-rate noise. These are concrete encoder fine-tuning / relabeling targets.

Dynamic- k shows potential but is highly algorithm-dependent. A natural idea is to size k per call from the encoder’s confidence (top- p , relative-threshold, largest-gap, margin/entropy 2-tier). Whether it helps depends sharply on the rule: against the *honest integer* static- k baseline at matched budget, some implementations improve noticeably while others fall below it. On the **full** population the entropy 2-tier rule buys a small recall gain (about +0.006 to +0.012), narrowing to a tie on the large silver set; on the **routed** (low-confidence) slice the outcome is mixed across rules and across human and silver references. We therefore read dynamic sizing as a lever with demonstrated *potential* rather than a settled result—too algorithm-dependent here to claim it reliably beats a well-chosen static k , or that it fails to.

One candidate explanation for why the routed branch is the harder place to realize a consistent gain is a **selection effect** (a hypothesis argued from our data, not proved): a confidence-gated cascade routes on the very signal (margin/entropy) that per-call sizing relies on, so conditional on being routed that signal is partly spent. The correlation between the encoder’s confidence and true-label depth drops from $|\rho| \approx 0.51$ on the full set to ≈ 0.11 – 0.25 on the routed slice, leaving a more uniformly uncertain subset to size over; the gate has already made the one high-value dynamic move, collapsing to $k \approx 1$ on the confident majority. This fits CSC’s advantage coming from the *calibrated accept/escalate decision* and per-band coverage rather than from resizing the shortlist on the routed branch, but it leaves room for a well-designed sizing rule to help—so we present dynamic- k as an open direction. The complementary accuracy lever remains ranker recall, exactly what CSC’s per-band guarantee (§4) is built to control.

Table 3: Within-cascade sensitivity: how shortlist size moves cost inside the escalated tier, relative to the deployed top-5 operating point. The width / recall / cost trade-off is monotone and mild. The recall column is escalated-tier recall at the deployed operating point, not the full-gold recall@5 of Table 2; the accuracy column is relative to the deployed top-5 point (percentage points of conditional-picker accuracy). Only the deployed top-5 point was run; the top-3 and top-10 rows are *projected* from an input-token cost model (App. A.9), not separately measured.

Setting (within escalated tier)	Cost vs. deployed	Shortlist recall	Accuracy vs. deployed
No shortlisting (full 100+ list)	higher	100%	reference (full-list acc.)
Top-10	+45%	98%	+1%
Top-5 (deployed)	baseline	95%	baseline
Top-3	−27%	90%	−5%

9 COST AND ECONOMICS

We report **relative** cost against the full-LLM-prompt baseline (every call sent to the picker with the full list of all 100+ categories); absolute dollar and token figures are deliberately omitted for anonymity.

Headline. The dominant saving is the encoder gate. At the deployed operating point the confident encoder auto-accepts roughly 87% of calls and escalates only $\sim 13\%$ to the picker; the escalated fraction is set by the coverage level—the system’s single knob—so more cautious operating points route a larger share (up to $\sim 21\%$, i.e. $\sim 79\%$ auto-accepted) at higher cost. Shortlisting compounds the saving *inside* the escalated tier: replacing the full 100+-label prompt with a five-candidate one cuts each escalated call to $\approx 0.14\times$ the input tokens ($\approx 2.9\times$ cheaper per call; \$5, App. A.9). Stacked on the gate, this brings the deployed cascade to more than 90% below the cost of sending every call to the picker with the full label list.

Inside the escalated tier, resizing the shortlist moves cost only mildly and monotonically (Table 3): dropping to top-3 saves 27% but gives up 5 points of conditional accuracy, while widening to top-10 buys back barely a point at 45% more cost. Read against \$5, this is the crux. Shortlisting was adopted to save picker tokens, yet an unadapted prompt caps accuracy, and widening the list to recover it stays cheap against the gate but never reaches full-list accuracy—so the accuracy that shortlisting costs is not bought back by shortlist sizing. The one untested escape is prompt caching: a cached full-list prefix could in principle recover most of full-list accuracy at near-shortlist marginal cost.

At tens of millions of calls a year, this is what a reduction of more than 90% buys: the same adjudication is routine under the cascade but prohibitive if the LLM saw every call over the full list—the difference between an economically viable pipeline and one that cannot ship.

10 DISCUSSION AND DEPLOYMENT GUIDANCE

The case study yields concrete guidance. (i) *The gate is the money; the shortlist is the accuracy risk.* Spend calibration effort on the escalation decision (where CSC’s single knob and per-band guarantee pay off) and be wary of aggressive shortlisting on the picker. (ii) *Prefer definition-grounded candidates* when shortlisting is used; it is the best prompt lever, though bounded. (iii) *Invest in ranker recall first*—hard-negative fine-tuning targeted at the ~ 8 –19 confusable categories that dominate misses is the more reliable lever on the routed slice, while per-call dynamic sizing is a promising but algorithm-dependent add-on worth tuning case by case. (iv) Alternative pickers worth testing include classify-then-map (name the intent first, then map to a candidate), pairwise tournaments, and elimination-first prompting, all of which may help recover information or decision context available in the full-list condition. Data-centric tail strategies (e.g., LLM-driven tail augmentation) are complementary to CSC’s inference-time adjudication.

11 LIMITATIONS

The human-gold hard slice is small ($n \approx 96\text{--}128$; $\pm \sim 9\text{pp}$), so gold numbers are directional and several dynamic- k “wins” are within sampling noise; silver corroborates trends but is optimistic and, because one silver labeling vote may share the picker’s model family, a full-list silver reference is partly self-referential—which is why the headline conditional-accuracy numbers and the shortlist transfer gap are reported on human gold, and the silver routed slice is used only for the corroborating error diagnostics of §6. The study covers a single operator, taxonomy, and language, and one LLM family as the picker. End-to-end picker-accuracy@ k is measured at the deployed operating points, not a full sweep. Prompt-caching feasibility is unconfirmed. Because the evaluated operating point is the fixed- (τ, k) special case, claims that adaptive conformal sets *dominate* fixed- (τ, k) at matched cost are stated as framework properties and design-time expectations, not as a head-to-head empirical result in this study. Finally, the conformal per-band coverage guarantee is a property of the construction; realized per-band coverage on test is reported as validation of the design, and we caution that the guarantee is only as good as calibration-set exchangeability under production drift. Anticipated reviewer concerns—“combination of known components” (answered by the subsumption argument and the per-band coverage guarantee, neither of which any single prior work provides), “why not a learned router” (learning-to-defer is a natural alternative to our calibrated gate, which we discuss but do not evaluate here), and “is the guarantee vacuous” (we report realized per-band coverage and set-size distributions)—are addressed in the appendix.

12 CONCLUSION

A single calibrated object—the encoder’s conformal confusion set—can drive both halves of a hybrid classifier: it accepts confident calls without ever invoking the LLM, and hands ambiguous ones exactly the labels the encoder cannot separate, with a per-frequency-band recall guarantee. But the shortlisted prompt is a prompt in its own right: one written for the full taxonomy transfers poorly when reused unchanged, scoring below the full list on human-gold routed calls even when the true label is present. Tuned for the shortlisted regime, it substantially improves conditional selection on paired human-gold calls ($42.7\% \rightarrow 63.5\%$); beyond that, the complementary ceiling is shortlist recall, governed by ranker quality rather than the picker prompt or any per-call sizing rule. For budget-constrained categorization over large, long-tailed taxonomies, the lesson is to calibrate the escalation decision well, ground the candidates you do show, and spend the rest of the budget on recall.

REPRODUCIBILITY STATEMENT

The evaluation protocol (splits, held-out human-gold set, the accuracy@ k given true-in-top- k metric) is specified in §3; the conformal construction, per-band scores, and budgeted policy in §4; the frozen validation-authored prompt assets (definitions, exclusion rules) and five-arm methodology in §7; and the recall/dynamic- k analysis, including the honest integer-static baseline, in §8. Anonymized code, prompt templates, and per-band coverage/recall tables are included in the supplementary material. Data are real customer-call summaries and cannot be released; we describe all processing steps and provide the category-level statistics needed to interpret the results.

ETHICS STATEMENT

The data are human-subject customer-care call transcripts containing personally identifiable information. All records are de-identified and access-controlled; we report only genericized category names and aggregate metrics, and no raw transcripts or operator identifiers. The per-band coverage guarantee has a fairness dimension: it explicitly protects rare categories from being silently dropped from shortlists, which matters when categories correspond to vulnerable-customer intents. No individual-level decisions are made from model outputs in this study.

AI USE STATEMENT

Large language models are both the object of study (the picker) and part of the labeling/definition-synthesis pipeline, as described in §3 and §7. LLM assistance was also used in drafting and editing

this manuscript; all technical claims, experiments, and numbers were produced and verified by the authors.

REFERENCES

- Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4): 494–591, 2023.
- Aabesh Bhattacharyya, Tiffany Ding, and Rina Foygel Barber. Conformal prediction with macro-coverage guarantees. *arXiv preprint arXiv:2606.28598*, 2026.
- Dylan Bouchard. Is escalation worth it? a decision-theoretic characterization of LLM cascades. *arXiv preprint arXiv:2605.06350*, 2026.
- Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.
- Zhu Cheng, Wen Zhang, Chih-Chi Chou, You-Yi Jau, Archita Pathak, Peng Gao, and Umit Batur. E-commerce product categorization with an LLM-based dual-expert classification paradigm. In *CustomNLP4U Workshop at EMNLP*, 2024.
- Hamvir Dev, Cijo George, Jeevesh Nandan, Anup Pattnaik, and Sasanka Vutla. Scalable and cost-effective high-cardinality classification with LLMs via multi-view label representations and retrieval augmentation. In *Proceedings of EMNLP 2025 (Industry Track)*, 2025.
- Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. *arXiv preprint arXiv:2507.06867*, 2025.
- Yifan Dou, Shikan Lian, and Shibo Li. Conformal cascade: Distribution-free accuracy guarantees for multi-tier LLM inference. *arXiv preprint arXiv:2607.25018*, 2026.
- Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. *arXiv:2307.02764*.
- Rishik Kondadadi and John E. Ortega. Learning to defer for adaptive model selection in clinical text classification. *arXiv preprint arXiv:2604.13285*, 2026.
- Varun Kotte. UCCI: Calibrated uncertainty for cost-optimal LLM cascade routing. *arXiv preprint arXiv:2605.18796*, 2026.
- Adrien Le Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. *arXiv preprint arXiv:2411.02988*, 2024.
- Zhenyi Lu, Jie Tian, Wei Wei, Xiaoye Qu, Yu Cheng, Wenfeng Xie, and Dangyang Chen. Mitigating boundary ambiguity and inherent bias for text classification in the era of LLMs. *arXiv preprint arXiv:2406.07001*, 2024.
- Shreya Shankar, Sepanta Zeighami, and Aditya Parameswaran. Task cascades for efficient unstructured data processing. *Proceedings of the ACM on Management of Data (PACMMOD)*, 2026. *arXiv:2601.05536*.
- Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior recalibration for imbalanced datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Antonios Valkanias, Soumyasundar Pal, Pavel Rumiantsev, Yingxue Zhang, and Mark Coates. C3PO: Optimized LLM cascades with probabilistic cost constraints. *arXiv preprint arXiv:2511.07396*, 2025.
- Nathan Vandemoortele, Bram Steenwinckel, Femke Ongena, and Sofie Van Hoecke. From haystack to needle: Label space reduction for zero-shot classification. *arXiv preprint arXiv:2502.08436*, 2025.

Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolò Dalmaso, Natraj Raman, and Sumitra Ganesh. Prune ’n predict (CROQ): Optimizing LLM decision-making with conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

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

Yiding Wang, Kai Chen, Haisheng Tan, and Kun Guo. Tabi: An efficient multi-level inference system for large language models. In *Proceedings of the Eighteenth European Conference on Computer Systems (EuroSys)*, 2023.

Ziji Zhang, Michael Yang, Zhiyu Chen, Yingying Zhuang, Shu-Ting Pi, Qun Liu, Rajashekar Maragoud, Vy Nguyen, and Anurag Beniwal. REIC: RAG-enhanced intent classification at scale. *Proceedings of EMNLP 2025 (Industry Track)*, 2025. arXiv:2506.00210.

Yaxin Zhu and Hamed Zamani. ICXML: An in-context learning framework for zero-shot extreme multi-label classification. *arXiv preprint arXiv:2311.09649*, 2023.

A APPENDIX

The main-text Related Work (§2) compresses two dozen references into four paragraphs. This appendix expands that discussion, adds the concrete mechanism and a headline result for each work, and states the single sharpest way each differs from the Conformal Shortlist Cascade (CSC). We keep the same organization: the “whether to escalate” axis (§A.1), the “what to expose” axis (§A.2), the direction of the set-size $\leftrightarrow$ accuracy relationship (§A.3), and the long-tailed calibration machinery CSC builds on (§A.4). Table A.1 summarizes the whole field against CSC on five capabilities. The recurring pattern is that prior work owns exactly one of CSC’s two coupled decisions: cascade work gates escalation while holding the LLM’s label space fixed, and shortlist work restricts the label space while calling the LLM on every input.

A.1 CASCADES AND ROUTING: THE “WHETHER TO ESCALATE” AXIS

This line decides *whether* to send a call to the expensive model; the label space the LLM ultimately sees is never a control variable.

Confidence-routed serving systems. FrugalGPT (Chen et al., 2023) queries a sequence of LLM APIs in cost order and accepts the first answer whose learned (DistilBERT-based) reliability score clears a per-stage threshold; the API list and thresholds are chosen jointly by constrained optimization. It matches GPT-4 accuracy on the HEADLINES task at up to 98.3% lower cost. Tabi (Wang et al., 2023) likewise routes by calibrated confidence in a multi-level serving system. Task Cascades (Shankar et al., 2026) carry the same escalation logic into unstructured-data-processing pipelines, ordering cheap operators ahead of expensive LLM calls so the LLM runs only when the cheaper stages cannot resolve a record. In all three, every model that runs sees the same task and answer space—routing chooses *which* model answers, never *what* it is asked to choose among.

Decision-theoretic treatments of deferral. Jitkrittum et al. (2023) derive the Bayes-optimal defer rule for a two-model cascade: defer when model 2’s probability of being correct exceeds model 1’s by more than the cost, which is a contrast of the two confidences rather than a threshold on model 1 alone. Plain confidence thresholding is provably suboptimal under specialist downstream experts, label noise, or distribution shift, and their post-hoc estimators (Diff-01, Diff-Prob) recover much of the gap. Bouchard (2026) characterize the cost–quality Pareto frontier of threshold cascades over a pool of models, showing the achievable frontier is the envelope of all pairwise two-model cascades; candidate-set restriction is not among its decision variables. Both keep the LLM’s task fixed and reason only about the escalation probability.

Budgeted and learned gates. UCCI (Kotte, 2026) isotonicly calibrates a token-margin uncertainty into a per-query error probability (ECE 0.12 $\rightarrow$ 0.03) and proves that thresholding the calibrated probability is cost-optimal under a cost-ordering and a routing-invariant-accuracy assumption, cutting cost 31% at fixed micro-F1. C3PO (Valkanis et al., 2025) tunes per-stage exit thresholds

from agreement with the most powerful model rather than gold labels—under 1% of the labels supervised baselines need—and uses a conformal quantile check to bound $\Pr(\text{cost} > C^*) \leq \alpha$; it reaches within 2% of the top model’s accuracy at under 20% of its cost. L2D (Kondadadi & Ortega, 2026) trains a logistic deferral head on encoder uncertainty plus text features to route clinical-text inputs between a fine-tuned BERT and an LLM, hitting F1 0.928 at 7% LLM calls. Across all three, the downstream LLM still receives the full classification task; C3PO’s conformal step bounds *total cost*, not a per-call candidate set, and none makes the exposed label space a decision variable. A learned gate of this kind is the natural alternative to our calibrated gate; we discuss it as future work rather than evaluate it here (App. A.10).

A.2 CANDIDATE-SPACE RESTRICTION: THE “WHAT TO EXPOSE” AXIS

This line makes LLM classification over large label spaces feasible by shortlisting candidates, but—with one partial exception—invokes the LLM on *every* input, with no cheap-model termination, no cost budget, and no per-class coverage guarantee.

Fine-tuned proposer plus LLM selector. The closest single mechanism is the Dual-Expert paradigm (Cheng et al., 2024): a fine-tuned XLM-R expert proposes top- k ($k=10$) candidates over a 2,000+ category catalog and an off-the-shelf Mixtral selects among them, lifting macro-F1 from 0.782 to 0.925. Notably, it is the one shortlisting work with a *gate*—an ad-hoc confidence threshold routes only $\sim 20\%$ of traffic through the LLM stage—but the gate is a hand-set heuristic with no coverage or recall guarantee, and the paper never asks whether a shorter list hurts the selector when the true label is present.

Retrieval-shortlisted classification. pattnaik2025 (Dev et al., 2025) retrieves candidates in the same contact-center domain as CSC using a multi-view label index (name, description, exemplars), grid-searching the top- k ($20 \leq k \leq 50$) subject to a $\leq 5\%$ retrieval-error budget, for up to 14.6% higher accuracy at 60–91% lower cost. REIC (Zhang et al., 2025) retrieves the top-10 (query, intent) pairs and has a LoRA-tuned Mistral score each by constrained decoding (F1 0.572 vs. 0.516 for a fine-tuned RoBERTa). ICXML (Zhu & Zamani, 2023) generates free-form candidates, maps them into a 131K–320K-label space by dense retrieval, and reranks with the LLM. Haystack-to-Needle (Vandemoortele et al., 2025) iteratively reduces the label list to $k = 2\text{--}5$ via a CatBoost distillation of LLM pseudo-labels, gaining +7.0% macro-F1 on average. All four call the LLM on every input; none gates on a cheap model, and none carries a per-frequency-band recall guarantee—pattnaik2025 develops an explicit cost analysis but still stops at a retrieval-error target rather than a per-class coverage floor.

A.3 THE SET-SIZE $\leftrightarrow$ ACCURACY RELATIONSHIP, AND WHY OUR SIGN DIFFERS

Several of these works report that *fewer* candidates *help* the LLM, which is the opposite sign from our shortlist penalty.

Lu et al. (2024) document that LLM classification accuracy collapses as the option count grows—94.3% at 2 options down to 32.5% at 60 on gpt-3.5-turbo—and attribute it to boundary ambiguity among near-synonym classes and to position/token bias, none of which longer context windows or scale remove; their option-reduction plus pairwise-contrastive prompt lifts accuracy from 40.2% to 62.0%. CROQ (Vishwakarma et al., 2025) conformally prunes the option set and re-prompts the *same* LLM on the surviving members, improving accuracy by 4–7 points and formalizing a monotone-accuracy assumption (fewer options, higher accuracy). Haystack (Vandemoortele et al., 2025) and the top- k ablation of pattnaik2025 (Dev et al., 2025) report the same direction.

These findings are consistent with ours once the conditioning is made explicit. Prior comparisons contrast a full list against a reduced list *without* conditioning on the true label being present, so their “gains” bundle a recall or attention-dilution benefit—removing distractors that the model would otherwise lose to—with any change in the selection task itself. CROQ’s benefit, in particular, is contingent on the coverage event that the pruned set still contains the answer, and its starting point is a small (≤ 15) multiple-choice set. CSC instead scores *conditioned on the true label being in the shortlist* (accuracy@k given true-in-top-k) and starts from a strong full-list baseline over 100+ classes. Under that leakage-controlled conditional metric on human-gold labels, restricting a strong

picker from the full list to five near-twin candidates with an unadapted prompt *drops* conditional accuracy (52.1% $\rightarrow$ 42.7%) even though the true label is present over 90% of the time. The prior sign and ours therefore measure different contrasts: distractor removal helps when it rescues recall, whereas narrowing a strong model to a few near-twins, holding recall fixed, changes the selection task the prompt must handle and can hurt when the prompt is not adapted for it. Documenting that second effect inside a cost-gated, long-tailed cascade is our contribution, not a claim to have discovered option-count sensitivity.

A.4 LONG-TAILED CALIBRATION AND CONFORMAL SETS

This machinery gives CSC tail-valid shortlists; none of it drives escalation or constructs an LLM candidate set.

Long-tailed conformal prediction. Standard conformal prediction hits marginal coverage but silently under-covers rare classes, while classwise conformal covers each class at the cost of enormous sets. Ding et al. (2025) resolve this with a prevalence-adjusted softmax score $s(x, y) = -\hat{p}(y | x)/\hat{p}(y)$ (with $\hat{p}(y)$ the marginal label prevalence), the size-optimal score for macro-coverage: on PI@ntNet-300K it cuts the number of under-covered species from 421 to 180 at far smaller sets than classwise. Bhattacharyya et al. (2026) generalize this to label-weighted conformal prediction with a finite-sample macro-coverage bound $\text{MacroCov} \geq 1 - \alpha - \max_k w(k)/N_k$ and an α' -correction that restores exact $1 - \alpha$, reaching the same guarantee at set size 2.4 where classwise needs 58.1. CSC adopts both—the prevalence-adjusted score and the label-weighted construction—but with *frequency bands* as the groups and independent per-band recall floors, where these papers target a single weighted average (macro-coverage) and stop at the prediction set.

Calibration under many and imbalanced classes. Le Coz et al. (2024) recast many-class confidence calibration as one balanced binary problem (Top-versus-All), driving ImageNet-21K ECE from 12.3% to 0.17% and avoiding the overfitting and class-flipping of vector/Dirichlet and one-versus-all scaling. Tian et al. (2020) give the Bayesian justification for prevalence adjustment under label-prior shift—the optimal target classifier scales $\hat{p}(y | x)$ by $p_t(y)/p_s(y)$ —and stabilize the naive correction by KL-interpolating $P_f^* \propto P_d^{1-\lambda} P_r^\lambda$ with a single tuned λ . Both sharpen the probabilities CSC’s conformal score consumes, but each emits a single calibrated decision or probability vector, not a set with a coverage guarantee, and neither addresses the recall of a rare class inside a shortlist.

A.5 CLOSEST NEIGHBORS AND POSITION

Two works sit directly adjacent to CSC and, between them, motivate it. Conformal Cascade (Dou et al., 2026) replaces the confidence threshold in a multi-tier cascade with a per-tier conformal set built from self-consistency frequencies and defers purely on set size ($|C|=1$ accept, $|C|>1$ escalate), with a clean cost model $\mathbb{E}[\text{cost}] = c_1 + c_2 \Pr[|C_1(X)| \neq 1]$ and up to 53% cost reduction at -5.3 points. Its coverage guarantee is marginal, not per-class, and—decisively—the escalated tier discards the set and re-scores the full answer space from scratch: the conformal set drives the *gate* but is never forwarded as a shortlist. CROQ (Vishwakarma et al., 2025) is the mirror image: it *passes* a conformal set to the LLM as its option list but has no cheap-model tier, so every query incurs an LLM call, and it carries no cost budget or long-tail structure.

CSC is precisely the union of these two neighbors—a conformal set that *both* gates escalation *and* becomes the LLM’s candidate shortlist—plus a per-frequency-band recall guarantee that neither provides, instantiated and cost-measured in a deployed system over 100+ long-tailed classes. To our knowledge no prior work occupies the intersection of a selective gate that terminates easy calls, adaptive restriction of the exposed label space, a supervised encoder base, explicit per-class long-tail analysis, and a measured cost budget at production scale (Table A.1).

The remaining subsections support the claims the main text defers here. §A.6 gives the conformal construction and the proof of the per-band coverage guarantee (answering “is the guarantee vacuous” and, with §4’s subsumption argument, “combination of known components”). §A.7 gives the per-arm picker prompts and the definition-synthesis procedure. §A.8 reports realized shortlist recall and the coverage/set-size quantities. §A.9 gives the relative-cost tables. §A.10 gives encoder train-

Table A.1: Positioning of prior work against CSC on five capabilities. *Gate*: a cheap model can terminate without invoking the LLM. *Set is control var.*: the label subset exposed to the LLM is a per-call decision variable. *Set $\rightarrow$ LLM*: a conformal (or pruned) set is forwarded as the LLM’s candidate list. *Cost budget*: an explicit inference-cost budget or bound. *Long-tail cov.*: a per-class or per-band coverage/recall guarantee. $\checkmark$ marks a full capability; (p) marks partial or ad-hoc; — marks absent. CSC is the only row with all five.

Work	Gate	Set is control var.	Set $\rightarrow$ LLM	Cost budget	Long-tail cov.
FrugalGPT (Chen et al., 2023)	$\checkmark$	—	—	$\checkmark$	—
Tabi (Wang et al., 2023)	$\checkmark$	—	—	$\checkmark$	—
Deferral (Jitkrittum et al., 2023)	$\checkmark$	—	—	(p)	—
Escalation (Bouchard, 2026)	$\checkmark$	—	—	$\checkmark$	—
UCCI (Kotte, 2026)	$\checkmark$	—	—	$\checkmark$	—
C3PO (Valkanias et al., 2025)	$\checkmark$	—	—	$\checkmark$	—
L2D (Kondadadi & Ortega, 2026)	$\checkmark$	—	—	(p)	—
Dual-Expert (Cheng et al., 2024)	(p)	$\checkmark$	$\checkmark$	(p)	—
Retrieval shortlist (Dev et al., 2025; Zhang et al., 2025; Zhu & Zamani, 2023; Vandemoortele et al., 2025)	—	$\checkmark$	$\checkmark$	(p)	—
CROQ (Vishwakarma et al., 2025)	—	$\checkmark$	$\checkmark$	—	—
Conformal Cascade (Dou et al., 2026)	$\checkmark$	—	—	$\checkmark$	—
Long-tail conformal (Ding et al., 2025; Bhattacharyya et al., 2026)	—	—	—	—	$\checkmark$
CSC (ours)	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$

ing/HPO details and discusses the learned-router (learning-to-defer) alternative (“why not a learned router”) as future work.

A.6 CONFORMAL CONSTRUCTION AND THE PER-BAND COVERAGE GUARANTEE

Construction. Let the encoder emit $\hat{p}(y \mid x)$ over the K categories (with x the input call and y a candidate label). We reserve a calibration split (the held-out silver validation split, $n_{\text{cal}} \approx 12.7k$; §A.10) and, following prevalence-adjusted / macro-coverage conformal scoring (Ding et al., 2025, which interpolates between a single marginal threshold and per-class-conditional thresholds), form a nonconformity score $s(x, y)$ from $\hat{p}(y \mid x)$. Partition the label space $\mathcal{Y}$ into m class-frequency bands $\{B_1, \dots, B_m\}$ by training prevalence. For band b with n_b calibration points whose true label lies in b , set the band threshold $\hat{q}_b$ to the $\lceil (n_b + 1)(1 - \alpha) \rceil$ -th smallest calibration score in that band. The confusion set is

$$C_\alpha(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}_{b(y)}\}, \quad (2)$$

where $b(y)$ is the band of label y . If $|C_\alpha(x)| = 1$ the encoder’s label is accepted; otherwise $C_\alpha(x)$ is escalated as the shortlist.

Guarantee. Calibrating a separate threshold within each band makes C_α an instance of *Mondrian* (class-conditional) split-conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023). The finite-sample coverage guarantee is therefore the standard conformal one, applied within a band: under exchangeability of the calibration and test scores inside band b , a test call whose true label y^* lies in band b satisfies

$$\Pr[y^* \in C_\alpha(x) \mid y^* \in B_b] \geq 1 - \alpha$$

(with the matching upper bound $1 - \alpha + \frac{1}{n_b + 1}$ when scores are almost surely distinct). We do not reprove this result; the quantile/exchangeability argument is given in Vovk et al. (2005) and, in tutorial form, Angelopoulos & Bates (2023), and its long-tailed / per-class instantiation in Ding et al. (2025); Bhattacharyya et al. (2026). Marginal coverage $\geq 1 - \alpha$ then follows as a prevalence-weighted mixture over bands, but per-band control is strictly stronger in the long tail: a marginally valid set can meet $1 - \alpha$ overall while systematically undercovering rare bands, which is what the per-band construction rules out.

Cost trade and what the deployed system evaluates. Two caveats bound the guarantee. First, the operational cap k_{max} on $|C_\alpha(x)|$ can drop the true label and void the $\geq 1 - \alpha$ guarantee—an explicit guarantee-for-cost trade. Second, and importantly for reading the numbers below: the operating point evaluated in the main paper is the *fixed*-(τ, k) *special case* of the budgeted policy (a top-5 shortlist at a margin gate τ on $\gamma = p_1 - p_2$; §A.9), not a separately tuned conformal sweep. The per-band construction above is the framework this point sits within; realized coverage is reported in §A.8 as *validation of the construction*, consistent with §4 and §11.

A.7 PER-ARM PICKER PROMPTS AND THE DEFINITION-SYNTHESIS PROCEDURE

The five-arm study (§7, Table 1) layers incremental, leakage-free changes on a common harness. All arms call the same picker (claude-sonnet-4-6, temperature 0, max_tokens= 1536) on the low-confidence slice routed by the production gate (encoder margin $\gamma < 0.93$) and emit an XML `<reasoning>/<category>` block so a single parser reads every arm. Arm A reproduces the deployed instruction content; Arm D additionally grounds each candidate in a synthesized definition.

Arm A (baseline, deployed instruction content).

I have a list of possible categories:
{enumerated_categories}

Here is the call information:
"{call_summary}"

Identify the best matching category from the provided list based specifically on the customer's initial intent for calling. Focus only on why the customer originally contacted support, ignoring final resolutions or transfer destinations. If multiple categories apply equally, choose the one that best represents the primary aspect of the customer's concern.

If *none* of the provided categories are a good fit, answer 'Other' for the best category and propose new category.

(In the deployed pipeline the trailing instruction is a plain-text "output only the category name"; the arm harness restores the original XML output block so parse_xml_response can read a reasoning trace.)

Arm D (definition-grounded shortlist prompt). Each of the 5 candidates is rendered as its name plus a 1–2 sentence "typical call" definition drawn from the frozen asset category_core_boundary.parquet:

Here is the call information:
"{call_summary}"

These 5 categories were pre-selected by an upstream model; usually the best answer is among them:
{enumerated_categories} % each: "i. {name}\n Typical: {definition}"

Classify the call into the single best category based specifically on the customer's initial intent for calling. Focus only on why the customer originally contacted support, ignoring final resolutions or transfer destinations.

Think step by step inside a `<reasoning>` block:

1. State the single fact about this call that most distinguishes among these 5 categories.
2. Rank all 5 categories from best to worst match, one line each, citing that fact to justify each candidate's position.

Then choose the top-ranked category as your answer. Choose 'Other' ONLY if you can state why each of the 5 categories is wrong.

Arms B, C, and E interpolate between these (B: calibrated framing + a justified "Other"; C: contrastive best-to-worst ranking; E: adds targeted pairwise exclusion rules on top of D).

Definition synthesis (CORE). The per-category definitions are built offline and leakage-free. For each category we sample $N_{\text{TARGET}} = 30$ grounding calls (and $N_{\text{SIBLING}} = 5$ contrast calls from sibling categories) *from the silver validation split only*, under a fixed random seed, and synthesize

Table A.2: Realized shortlist recall (fraction of calls whose true label is in the top-5), by split and by reference. Silver-referenced hit rates are measured on the full splits; the human-referenced rates separate the full golden set from the low-confidence *hard* slice that is actually routed.

Reference	Slice	Recall@5
Silver consensus	Golden (hit@5 / hit@10)	93.1% / 96.2%
Silver consensus	Validation	93.8%
Silver consensus	Test	94.7%
Human gold	Full golden set ($n \approx 900$)	93.1%
Human gold	Routed hard slice ($n \approx 96$)	80.0%

Table A.3: Within escalated-tier width/recall/cost trade-off, relative to the deployed top-5 point. Recall is escalated-tier shortlist recall; accuracy is percentage points of conditional-picker accuracy relative to top-5. Cost is computed from **input prompt length** (input tokens), which the picker cost is dominated by; the per-candidate contribution scales with the number of shortlisted category definitions rendered in the prompt. Only the deployed top-5 point was run; the off-baseline rows (top-3, top-10) are *projected* from this input-token model, not separately measured.

Setting	Cost vs. top-5	Shortlist recall	Accuracy vs. top-5
No shortlist (full list)	higher	100%	full-list acc.
Top-10	+45%	98%	+1%
Top-5 (deployed)	baseline	95%	baseline
Top-3	-27%	90%	-5%

a positively phrased “CORE” (“typical call” plus defining criteria) together with sibling-contrast “boundary” text (used by Arm E). These are written once to `category_core_boundary.parquet` (columns `core`, `boundary`, `boundary_by_sibling`, `sibling_categories`, `n_examples`), frozen, and applied unchanged to test, exactly as a promoted production prompt would behave.¹

A.8 REALIZED SHORTLIST RECALL, COVERAGE, AND SET-SIZE DISTRIBUTIONS

Because the evaluated operating point is a fixed top-5 shortlist (§A.6), the escalated-tier confusion-set size is constant at 5 by construction; realized *recall* of the true label into that shortlist is the coverage quantity we can measure, reported in Table A.2. The deployed operating point itself was selected by a grid search over the margin threshold τ and shortlist size k : we took the point past which recall falls off sharply, trading a small amount of picker accuracy for coverage, which is how the top-5 shortlist at the deployed margin gate was fixed. What that grid search does *not* produce is the per-frequency-band coverage breakdown or the conformal set-size distribution $|C_\alpha(x)|$ under a swept α ; those are properties of the conformal framework (§A.6) that the fixed- (τ, k) deployment does not instantiate, and we leave the full α -sweep to future work.

A.9 RELATIVE COST AND THE WIDTH/RECALL/COST TRADE-OFF

All costs are **relative** (absolute token and dollar figures are omitted for anonymity). Two gate-level quantities underlie the §9 headline. Routing only the escalated tier cuts picker cost by roughly $5\text{--}8\times$ across the deployed operating range ($\sim 4.8\times$ at the cautious 79%-accept end, rising toward $\sim 8\times$ at the deployed 87% point), and the Arm-D shortlist is $\approx 2.9\times$ cheaper than the full-list prompt. The token-level breakdown: the full-list prompt uses $\approx 8.0k$ input tokens per call against the Arm-D shortlist’s $\approx 1.1k$ (about $0.14\times$ the input), at the cost of a longer reasoning output ($\approx 0.35k$ vs. $\approx 0.07k$ output tokens). Table A.3 gives the within-tier width/recall/cost trade-off referenced by Table 3; it is monotone and mild.

¹The synthesis notebook targets a 2–3 sentence CORE; the deployed definitions render as 1–2 sentences after truncation. We report the as-rendered length.

Table A.4: DeBERTa fine-tuning hyperparameters (Bayesian HPO winners). Objective is silver-validation accuracy. Large is the deployed encoder; base is reported for comparison.

Hyperparameter	DeBERTa-large (deployed)	DeBERTa-base
Learning rate	2.10×10^{-5}	2.62×10^{-5}
Batch size	32	32
Epochs	5	5
Warmup ratio	0.110	0.118
Weight decay	4.06×10^{-3}	2.40×10^{-2}
Dropout	0.137	0.149
Max sequence length	512	512
Early-stop patience	3	3
Seed	42	42
Validation accuracy (objective)	0.830	0.817
Validation macro-F1	0.824	—
Human-gold top-1 accuracy	0.732	—

A.10 ENCODER TRAINING, HPO, AND THE LEARNED-ROUTER ALTERNATIVE

Data. The silver set has $\sim 63k$ calls over the K categories with high panel consensus, split stratified 60/20/20 (seed 42) into $\sim 38k$ / $\sim 12.7k$ / $\sim 12.7k$ train/validation/test. The truth-bearing human-gold evaluation set (~ 900 calls) is held out of training entirely.

Encoder and HPO. The encoder is microsoft/deberta-v3-large (a K -way head). The Bayesian-search HPO winner (Table A.4) is used for all reported results.

Learned-router (learning-to-defer) alternative. A learned router—a learning-to-defer head trained to predict accept-vs-defer, in the sense of Kondadadi & Ortega (2026)—is a natural alternative to our calibrated margin gate. We do not evaluate it here and leave the head-to-head comparison at matched escalation budget to future work.