

Cost-Efficient Intent Classification via Semantic-Candidate Adaptation and Selective Capability Routing

Jiawei Sun and Yongsheng Sang*

College of Computer Science, Sichuan University

Chengdu, China

2025223045164@stu.scu.edu.cn, sangys@scu.edu.cn

Abstract

Intent classification deployments trade accuracy against retraining effort and inference cost. We study frozen generative reranking and budgeted large-model escalation, introducing Semantic-Candidate Adaptation (SCA): BGE top-5 retrieval followed by joint reranking with one representative training example per candidate. On CLINC150, aligned examples raise Qwen2.5-3B-Instruct from 68.18% label-only accuracy to 76.13% and Gemma-2-2B-it from 66.02% to 76.15%, without LLM parameter updates. CLINC150 wrong-example controls support the role of correct alignment, while BANKING77 and HWU64 provide transfer evidence. Broader comparisons clarify the deployment regimes: direct centroid retrieval reaches 80.71%, while a task-trained MPNet Energy baseline reaches 90.05%; always-on SCA is therefore not a universal replacement for direct or trained classifiers. A validation-selected low-margin SCA fallback provides only a marginal gain at an 8.62% realized test call rate. For real Qwen3.7-Max escalation on CLINC150, routing the lowest BGE top-1 retrieval scores reaches 82.11–82.89% accuracy at a 10% budget and 87.33–88.07% at 20%, with OOS F1 of 70.00–70.13% and 87.47–87.70%, respectively. Learned and token-confidence routing remain alternatives, but no policy dominates every evaluated point.

1 Introduction

Intent classification is a common front-end component of task-oriented dialogue, customer-service, and enterprise automation systems. Recent large language models (LLMs) can perform intent classification with strong zero-shot or few-shot behavior, but applying a large model to every request is often unnecessary and expensive. In many production settings, most requests are simple, repetitive, and in-scope; only a smaller fraction of ambiguous, out-of-scope (OOS), or fine-grained boundary

cases require stronger reasoning. This motivates hybrid inference systems that combine small and large models under a cost budget (Ding et al., 2024; Chen et al., 2024; Ong et al., 2025).

The direct use of small generative models for intent classification remains brittle. When the full intent space is large, directly prompting a small model with all labels can lead to parsing failures, label confusions, and poor OOS behavior. Our experiments on CLINC150 (Larson et al., 2019) show that direct full-label prompting gives only 51.13% accuracy for Qwen2.5-3B-Instruct and 54.24% for Gemma-2-2B-it. This is not because the correct label is impossible to retrieve: with a BGE embedding retriever, the gold label appears in the top-5 candidate set for 92.16% of CLINC150 test examples. The main bottleneck is therefore not only candidate recall, but whether the small generative model can understand fine-grained intent boundaries within a compact candidate set.

We separate two questions that are often conflated. First, can representative evidence improve a frozen generative reranker? Semantic-Candidate Adaptation (SCA) applies BGE top-5 retrieval and attaches one representative training example per candidate, changing only the prompt. Second, which deployment policy should decide between a local prediction and a stronger API model? We compare a simple retrieval-score rule, small-model token confidence, and a learned multi-signal router using retrieval ambiguity, linguistic complexity, and optional failure-graph features.

This paper is written from a deployment perspective. The experiments identify distinct operating regimes rather than one universal ranking. A task-trained encoder-head system is preferable when supervised retraining is acceptable; direct centroid retrieval is attractive when it is already reliable; SCA improves candidate discrimination when a frozen generative backbone is required; and large-model escalation trades external cost for stronger

*Corresponding author.

OOS handling. We retain SetFit (Tunstall et al., 2022) as a trained reference rather than claiming that frozen generative models replace compact classifiers.

Our contributions are:

- We identify a label-boundary bottleneck in small generative intent classification: high candidate recall does not guarantee high reranking accuracy.
- We propose SCA, which uses one aligned representative example per candidate to improve frozen generative reranking; matched wrong-example controls and cross-dataset results isolate and test this effect.
- We compare frozen reranking with direct centroid retrieval, a frozen cross-encoder, and task-trained OOD baselines, including candidate/example sensitivity and a selective-SCA analysis.
- We evaluate retrieval-score, token-confidence, and learned routing with the same real Qwen3.7-Max fallback, reporting complete OOS and cost metrics while treating failure-graph features as exploratory.

2 Related Work

Arora et al. (2024) highlight the strength of SetFit-style classifiers and study uncertainty-based routing for LLM intent detection. Chen and Wang (2025) also collaborate small and large models for few-shot intent detection, fine-tuning a task-specific small predictor with LLM-augmented data and invoking LLM refinement from its uncertainty. We instead keep the small generative LLM parameters fixed via SCA and compare retrieval-score, token-confidence, and learned routing under the same fallback, with cost and latency accounting.

Retrieval lets a model expose only a relevant slice of a large label space. Yu et al. (2021) predict from retrieved labeled examples and adapt to new domains by swapping the index without retraining, while Milios et al. (2023) retrieve demonstrations over large label spaces on CLINC150, BANKING77, and HWU64; a related strand scales intent classification to large, continually extended label sets (Wang et al., 2021). Most related to us, Zhang et al. (2025) retrieve query–intent pairs as candidate evidence and score the resulting intents with a customized LLM for large-scale customer service, and Liu et al. (2024) combine a fine-tuned smaller model with retrieval augmentation for multi-turn classification. Both adapt the underlying model by

fine-tuning and target proprietary multi-turn or multilingual settings; we instead share only the retrieve-then-rerank-with-evidence formulation while keeping the 2–3B generative model entirely frozen, isolating the effect of label-grounded examples with shifted- and random-wrong controls, and evaluating on public single-turn benchmarks.

Cost-aware hybrid inference sends easy inputs to a small model and hard ones to a large model. Ding et al. (2024) route by predicted query difficulty, Jitkrittum et al. (2023) analyze settings where confidence-only deferral is suboptimal, and FrugalGPT (Chen et al., 2024) and RouteLLM (Ong et al., 2025) learn general cost–quality routers. We test retrieval similarity, generated-token confidence, and a learned intent-specific router under the same fallback. For OOS detection, we compare with maximum-softmax-probability scoring (Hendrycks and Gimpel, 2017) and energy scoring (Liu et al., 2020), while distinguishing these controlled baselines from dedicated open-intent methods (Zhang et al., 2020, 2021; Zhou et al., 2023).

3 Method

3.1 Problem Formulation and System Overview

Given an input utterance x and an intent set $\mathcal{Y}$, the system predicts an intent label $y \in \mathcal{Y}$. We consider a small generative model S , such as Qwen2.5-3B-Instruct (Qwen Team et al., 2024) or Gemma-2-2B-it (Gemma Team et al., 2024), and a stronger large model L , here Qwen3.7-Max accessed through an API. The goal is to maximize classification quality while limiting the large-model call rate.

The framework has three components: (i) *candidate retrieval*, which retrieves a compact candidate set $C_k(x)$; (ii) *semantic reranking*, which asks the small generative model to choose among candidates with label-level evidence; and (iii) *residual routing*, which sends selected queries to the large model. Retrieval-score routing acts before generation and lets routed queries bypass the small model. Token-confidence routing and the learned router used in our original API runs act after small-model generation; the latter can additionally query an offline failure graph.

For each label y , we average the BGE embeddings of its training utterances to obtain centroid c_y . Candidate retrieval is:

$$C_k(x) = \text{TopK}_{y \in \mathcal{Y}} \text{sim}(e(x), c_y), \quad (1)$$

where $e(\cdot)$ is an embedding function and $k = 5$ in our main experiments. CLINC150 uses 151 centroids, including an OOS centroid formed from the 100 labeled OOS training utterances. The small model predicts:

$$\hat{y}_S = \arg \max_{y \in C_k(x)} p_S(y \mid x, C_k(x), s(y)), \quad (2)$$

where $s(y)$ is semantic evidence for candidate y , including the label name and a representative utterance from the training split.

The learned router outputs a failure score:

$$q(x) = \sigma(w^\top \phi(x, C_k(x))), \quad (3)$$

where ϕ contains complexity, retrieval ambiguity, and optional failure-graph features. We use $q(x)$ for ranking rather than calibrated probability estimation. Alternative policies rank by ascending BGE top-1 similarity or by ascending generated-token confidence. Given route budget r , let $g_r(x) \in \{0, 1\}$ indicate whether x is selected by a policy, and let $\hat{y}_L$ denote the large model’s prediction. The cascade prediction is:

$$\hat{y}(x) = \begin{cases} \hat{y}_L, & g_r(x) = 1, \\ \hat{y}_S, & g_r(x) = 0. \end{cases} \quad (4)$$

For a test set $\{(x_i, y_i)\}_{i=1}^n$, the realized route rate and cascade risk are:

$$\begin{aligned} B &= \frac{1}{n} \sum_{i=1}^n g_r(x_i), \\ R_{\text{cascade}} &= \frac{1}{n} \sum_{i=1}^n \mathbb{I}[\hat{y}(x_i) \neq y_i]. \end{aligned} \quad (5)$$

For a pre-generation router, with routed set $R_r = \{x : g_r(x) = 1\}$, the inference cost is:

$$\text{Cost} = (N - |R_r|) \cdot \text{Cost}_S + |R_r| \cdot \text{Cost}_L, \quad (6)$$

where N is the number of requests and $|R_r|/N = r$. Post-generation policies instead incur $N \text{Cost}_S + |R_r| \text{Cost}_L$. Since small-model inference is hosted locally, we separately report external API cost per 1K queries from routed large-model token usage and give measured or estimated latency with its execution order stated.

3.2 Semantic-Candidate Adaptation

Candidate retrieval. Since directly presenting all labels to a small model is difficult when the

intent set is large, we retrieve top-5 candidate intents before generation. We compare TF-IDF and BGE-style semantic retrieval (Xiao et al., 2024). On CLINC150 test, TF-IDF top-5 gives 81.42% candidate recall, while BGE top-5 gives 92.16%. On BANKING77 (Casanueva et al., 2020) and HWU64 (Liu et al., 2021), BGE top-5 recall is around 98.7%, so retrieval is not the dominant bottleneck with semantic embeddings. For CLINC150, oos is treated as the 151st label in the retrieval index; candidate-based methods can output oos only when it is retrieved into the top-5, with no separate “none of the above” option. Recall is computed over all test queries: BGE reaches 99.51% in-scope and 59.10% on oos, the latter upper-bounding small-model OOS accuracy.

Candidate evidence. Under SCA, each retrieved candidate receives lightweight semantic evidence, e.g., intent: spelling; example: what is the spelling of potato. For the main runs, we use one fixed in-label training utterance per label, taking the first occurrence in the prepared training file; the resampling check in Section 4.2 redraws this utterance from the same training split. This does not fine-tune the small model; it only changes the inference prompt. Intent labels are often underspecified or overlapping (e.g., calendar_update, todo_list_update, and reminder_update are close by name); boundaries become clearer once each label is grounded with a real utterance.

Negative controls. To check whether the SCA gain comes only from longer prompts or a more regular format, we run two controls: a *shifted-wrong* example, where each label receives an example from another label using a deterministic shift, and a *random-wrong* example, where each label receives a random example from an incorrect label. The complete CLINC150 label-only/aligned/misaligned comparison supports boundary grounding rather than prompt length; corresponding results on BANKING77 and HWU64 test whether the trend transfers.

Selective invocation. Direct centroid retrieval can outperform always-on generative reranking. We therefore also evaluate a selective variant that retains the centroid prediction except for queries with a small top-1/top-2 retrieval-score margin, where it invokes cached SCA. The margin threshold is fixed on validation for a target call rate and applied unchanged to test; this small-model fallback

is distinct from large-model API routing.

3.3 Selective Capability Routing

We compare three routing policies. *Low BGE top-1* routes the lowest retrieval-similarity scores and requires no router training or small-model generation for routed queries. *Token confidence* routes the lowest mean log-probability of generated label tokens under the original SCA prompt, excluding EOS. The *learned router* is logistic regression trained on validation labels indicating whether the small model is wrong, using retrieval, complexity, and optional failure-graph features.

Retrieval ambiguity. These features measure uncertainty in the retrieved candidate space: top-1 retrieval score, top-1/top-2 margin, score ratios, entropy over the top-5 scores, and OOS-related scores. They are computed before generation from retrieval alone.

Complexity. These features capture surface-level query properties such as length, token statistics, digits, question patterns, and rare-token indicators. They describe the demand imposed by the query rather than any model’s confidence.

Failure graph. These features summarize common confusions observed from both small models. Among 3,100 validation cases, 660 are wrong for both backbones, yielding 414 unique directed gold-to-wrong-label pairs, 227 of which occur at least twice. Features include candidate-neighborhood edge counts and densities and, after generation, predicted-label connectivity and cross-model agreement. Test labels are never used.

The submitted API runs used retrieval, complexity, and failure-graph features and therefore ran S before routing. Our new comparison changes only the routing score while keeping frozen SCA predictions, route budgets, the OOS-aware Qwen3.7-Max prompt, and decoding fixed. Because failure-graph removal yields small, mixed oracle changes (Section 4.8), we treat the graph as exploratory rather than a core source of routing gains.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on CLINC150, BANKING77, and HWU64. CLINC150 includes in-scope intents and OOS examples, making it the main testbed for routing. BANKING77 and HWU64 are

used to check whether SCA generalizes to other fine-grained intent datasets.

Models. The small generative models are Qwen2.5-3B-Instruct and Gemma-2-2B-it. The large fallback is Qwen3.7-Max, accessed through an API with OOS-aware prompting. Compact references use all-mpnet-base-v2 (Song et al., 2020) and the frozen 278M-parameter BAAI/bge-reranker-base checkpoint (Xiao et al., 2024).

Baselines. We compare direct full-label prompting, BGE top-5 label-only reranking, SCA, direct centroid retrieval, a local fixed-hyperparameter SetFit MPNet run, a task-trained MPNet linear head with Energy or MSP OOS scoring, a frozen BGE cross-encoder reranker, and three real routing policies at 10%/20% budgets. The SetFit row is a same-split reference, not a reproduction of Arora et al. (2024). Energy/MSP train the 150-way head on 15,000 in-scope utterances without OOS training utterances; labeled validation OOS data calibrate a threshold that is then frozen on test.

Inference and hardware. Generative decoding is deterministic (temperature 0, top_p=1, at most 16 generated tokens). Local SCA and cross-encoder throughput were measured on one Tesla V100-SXM2-16GB; reported cross-encoder times exclude the shared retrieval stage and use batch sizes 64/2/4 for BGE/Qwen/Gemma, so they are throughput references rather than batch-1 latency rankings.

Statistics. For the original CLINC150 SCA comparisons, we compute 95% CIs with 2,000 utterance-level bootstrap resamples. Added selective/routing comparisons use 10,000 utterance-level paired bootstrap resamples. We use exact paired McNemar tests and apply Holm correction when comparing the four backbone–budget settings as one family.

4.2 CLINC150 Main Results

Table 1 shows the main CLINC150 test results: direct full-label prompting is weak for both small models, BGE top-5 retrieval improves performance, and SCA (shown as BGE + 1 ex.) yields a further substantial gain.

The results have two implications. First, small generative models are highly sensitive to how the label space is exposed: under direct full-label prompting, neither model predicts OOS. Second, the SCA

Model	Method	Acc.	In-scope	OOS	Cand. recall
Qwen	Direct	51.13	62.49	0.00	–
Qwen	BGE labels	68.18	81.40	8.70	92.16
Qwen	BGE + 1 ex.	76.13	<u>89.16</u>	17.50	92.16
Gemma	Direct	54.24	66.29	0.00	–
Gemma	BGE labels	66.02	77.09	16.20	92.16
Gemma	BGE + 1 ex.	<u>76.15</u>	88.24	<u>21.70</u>	92.16
SetFit	Plain	86.13	95.87	42.30	–

Table 1: CLINC150 test results. Accuracy values are percentages. Qwen = Qwen2.5-3B-Instruct; Gemma = Gemma-2-2B-it; SetFit = SetFit MPNet. The Qwen and Gemma variants do not fine-tune the small generative model. Bold indicates the best result in each column of this table. Underlined values indicate the best result among frozen Qwen/Gemma variants.

gains over label-only reranking are statistically reliable: +7.95 points for Qwen (95% CI [7.09, 8.78]) and +10.13 for Gemma (95% CI [9.18, 11.07]), with paired McNemar tests indicating significance ($p < 0.001$). Across three resampling runs in which the representative example for each label is re-drawn from the training split under deterministic decoding, performance remains stable at 77.21 ± 0.82 for Qwen and 76.67 ± 0.46 for Gemma.

4.3 Operating-Regime Comparisons

Table 2 adds direct-retrieval and task-trained OOD references. Energy follows Liu et al. (2020); MSP uses the same frozen-encoder/head protocol. Energy and MSP outperform always-on SCA but require task-specific head training. The 150-centroid row removes the OOS prototype and calibrates its maximum-similarity threshold on validation, matching their OOS-supervision protocol.

Sensitivity. On CLINC150 validation, $k = 3$ versus $k = 5$ changes candidate recall from 97.97% to 98.71%, but Qwen/Gemma reranking accuracy from 88.87/88.19% to 87.61/86.97%. At $k = 5$, using 0, 1, and 4 examples gives 85.97/87.61/87.16% for Qwen and 81.48/86.97/85.71% for Gemma. Thus one example helps both backbones, whereas four provide no further gain over one (full table in Appendix B).

Frozen reranker baseline. On the same top-5 candidates, frozen bge-reranker-base reaches 57.51, 56.07, 56.13, and 59.89% overall accuracy with raw-label, humanized-label, structured-label, and structured-plus-example candidate text. The last construction receives the same candidate identities and representative utterances as SCA,

but scores each pair independently; it remains roughly 16.24–16.26 percentage points below the joint generative SCA reranking results for Qwen and Gemma, respectively. In this throughput setting, its 5.55 ms/query measurement is lower than Qwen/Gemma’s 130.7/128.8 ms/query; different batch sizes and inference paradigms preclude a strict one-to-one latency comparison. Appendix B reports all templates and integrity checks.

Direct and selective retrieval. Always-on SCA does not outperform 151-centroid top-1 accuracy (76.13/76.15% vs. 80.71%). Qwen/Gemma SCA fix 176/166 centroid errors but regress 428/417 correct predictions. With a low-margin threshold fixed at a 5% validation target, selective SCA invokes the small model on 8.62% of test examples and reaches 81.18/81.13% (net +26/+23; paired-bootstrap 95% CIs [+0.05,+0.89]/[+0.04,+0.80] points). At the 10% target the intervals include zero, and targets of at least 20% have non-positive net changes. This supports only a narrow selective fallback, not a general replacement for direct retrieval.

4.4 Cross-Dataset Generalization

Table 3 shows that SCA improves both small models on BANKING77 and HWU64.

SetFit remains stronger when task-specific training is allowed; our claim is not dominance over trained classifiers, but a low-maintenance adaptation path when fine-tuning or retraining is undesirable.

4.5 One-Example Control

Table 4 compares correct examples with wrong-example controls: the correct example consistently outperforms both, indicating that the gain is not only due to extra prompt tokens.

The controls suggest that SCA acts as boundary grounding: it helps the model decide what each retrieved label means in context.

4.6 Real API Routing

Table 5 compares three policies with the same frozen SCA predictions, route budgets, OOS-aware Qwen3.7-Max prompt, and deterministic decoding; only the routing score changes. The results are actual cascades, not oracle estimates.

At 10%/20%, Low BGE gives the highest accuracy in three of four settings and OOS F1 in all four. With cached 5%/15% points included, it leads accuracy at 5/8 and OOS F1 at 7/8 points;

Method	Parameter update	OOS train utterances	Acc.	In-scope	OOS P/R/F1	AUROC	IS→OOS FPR (%)
151-centroid top-1	None	100	80.71	93.91	99.07/21.30/35.06	–	0.04
150 centroids + threshold	None	0	86.93	93.31	92.82/58.20/71.54	96.32	1.00
MPNet MSP	Linear head	0	89.29	95.91	89.88/59.50/71.60	96.26	1.49
MPNet Energy	Linear head	0	90.05	96.13	93.72/62.70/75.13	97.86	0.93

Table 2: CLINC150 OOS baselines (percentages). The 150-centroid, MSP, and Energy thresholds use labeled validation OOS data but no OOS training utterances. The 151-centroid index instead includes the 100 labeled OOS training utterances, so its OOS metrics are not supervision-matched to the other rows.

Data	Model	Direct	BGE	SCA/SetFit	Gain	Recall
BANK	Qwen	50.26	74.45	<u>79.29</u>	+29.03	98.73
BANK	Gemma	52.50	62.08	<u>77.82</u>	+25.32	98.73
HWU	Qwen	69.52	74.16	<u>80.39</u>	+10.87	98.70
HWU	Gemma	66.36	70.72	<u>80.11</u>	+13.75	98.70
BANK	SetFit	–	–	89.51	–	–
HWU	SetFit	–	–	91.08	–	–

Table 3: Cross-dataset test results. BANK = BANK-ING77. BANKING77 and HWU64 have no OOS in the prepared split. SetFit rows report independent SetFit-MPNet baseline accuracy in the SCA/SetFit column; SetFit does not use retrieval or +1 examples. Bold indicates the best overall result in each dataset. Underlined values indicate the best result among frozen Qwen/Gemma variants.

Data	Model	Labels	Correct	Shifted	Random
CLINC	Qwen	68.18	76.13	72.69	72.04
CLINC	Gemma	66.02	76.15	70.45	71.42
BANK	Qwen	74.45	79.29	74.90	74.06
BANK	Gemma	62.08	77.82	72.60	70.65
HWU	Qwen	74.16	80.39	76.39	76.39
HWU	Gemma	70.72	80.11	75.00	76.21

Table 4: One-example negative controls. Labels denotes label-only reranking; Correct, Shifted, and Random denote one-example prompt variants.

no policy wins every point. Token confidence is informative for small-model error detection (AUROC 81.2%/78.2% for Qwen/Gemma), but neither it nor learned routing is consistently superior in the cascade. For the two Qwen learned-versus-confidence comparisons, Holm-adjusted McNemar p -values are both 0.137, so we interpret the differences descriptively. We do not report cascade-level AUROC because the final cascades emit discrete labels rather than a shared calibrated OOS score.

During post-submission audit, two Qwen API records contained transient transport failures and no model output but had been counted as errors. The corrected evaluation fills only these missing records with outputs from another run using the same model, prompt, and deterministic decoding;

all originally successful outputs remain unchanged, and no output selection is performed. This changes Qwen accuracy by +0.02/+0.04 points. At 10%/20%, corrected learned-router routed-large accuracy is 92.18%/89.09%, with 338,633/677,439 tokens as reported in Table 6.

4.7 Latency and Cost Analysis

We compute external API cost per 1K queries from the recorded input and output token counts using the DashScope price listed during the experiments for qwen3.7-max-2026-05-20: 12/36 CNY per million input/output tokens. Local inference counts as zero external cost; Table 6 reports the summed token usage.

Low BGE acts before generation, giving estimated Qwen/Gemma latency of 183.6/181.9 ms at 10% and 236.5/235.0 ms at 20% under $(1-r)t_S + rt_L$. Learned and token-confidence policies run SCA first, giving serial estimates 196.7/194.8 and 262.7/260.8 ms under $t_S + rt_L$. Here $t_L = 660.0$ ms is measured on a 100-sample OOS-free subset and is only a latency/cost reference. At 20%, external cost is about 1.49 CNY per 1K queries versus 7.40 for all-large calls.

4.8 Routing Analysis

The actual API caches are complete through 20%; results at 30/50/100% are oracle diagnostics only. In those diagnostics, every routed utterance is hypothetically corrected to its gold label; these values are theoretical upper bounds and do not represent real-model cascade outputs. Full policy curves and routed-subset composition are in Appendix A.

Failure-graph features do not add consistent oracle gain. Removing them changes learned-router oracle accuracy from 83.02/89.18 to 83.11/89.42 for Qwen and from 83.47/89.00 to 83.42/89.49 for Gemma at 10%/20%. We therefore retain the graph only as an exploratory error-structure analysis.

Small	Budget	Policy	Acc.	IS acc.	OOS P/R/F1	IS→OOS FPR (%)
Qwen	10%	Learned	82.24	89.80	94.88/48.20/63.93	0.58
Qwen	10%	Low BGE	82.89	89.18	97.50/54.60/70.00	0.31
Qwen	10%	Token conf.	81.38	90.38	91.29/40.90/56.49	0.87
Qwen	20%	Learned	87.00	90.38	92.88/71.80/80.99	1.22
Qwen	20%	Low BGE	88.07	89.64	95.07/81.00/87.47	0.93
Qwen	20%	Token conf.	86.13	91.44	90.54/62.20/73.74	1.44
Gemma	10%	Learned	82.42	89.04	93.76/52.60/67.39	0.78
Gemma	10%	Low BGE	82.11	88.20	97.68/54.70/70.13	0.29
Gemma	10%	Token conf.	82.20	89.31	94.36/50.20/65.54	0.67
Gemma	20%	Learned	86.45	89.84	91.75/71.20/80.18	1.42
Gemma	20%	Low BGE	87.33	88.67	95.20/81.30/87.70	0.91
Gemma	20%	Token conf.	85.76	90.27	91.48/65.50/76.34	1.36

Table 5: Real OOS-aware Qwen3.7-Max fallback on CLINC150 (percentages). Low BGE routes the lowest top-1 retrieval similarities; token confidence routes the lowest mean generated-label-token log-probabilities. Bold marks the best accuracy and OOS F1 per backbone/budget.

Policy	10%	20%
<i>Qwen small model</i>		
Learned	338,633	677,439
Low BGE	338,959	677,743
Token conf.	338,910	677,711
<i>Gemma small model</i>		
Learned	338,600	677,445
Low BGE	338,958	677,741
Token conf.	339,018	677,966

Table 6: Routed Qwen3.7-Max tokens at 10%/20% budgets. External cost, computed from these counts at the listed DashScope price for qwen3.7-max-2026-05-20, is CNY 0.74/1.49 per 1K queries. The 100-sample OOS-free all-large reference has 98.00% in-scope accuracy and costs CNY 7.40 per 1K queries.

4.9 Additional Analysis

At 10%, the learned router’s routed subsets contain 68.91% (Qwen) and 73.27% (Gemma) small-model errors, showing that it concentrates risk even though it is not uniformly the best policy. Table 9 further tests model participation under the oracle diagnostic. A pre-generation retrieval+complexity router reaches 83.11/89.42 (Qwen) and 83.42/89.49 (Gemma) at 10%/20%; source-to-target rows evaluate Qwen/Gemma transfer without target-model error labels. These two-backbone oracle results do not establish live-API transfer to unseen models.

5 Conclusion

We studied frozen candidate reranking and budgeted escalation for intent classification. Correctly aligned examples consistently improve frozen 2–3B rerankers, but direct retrieval and task-trained encoder-head systems are stronger in

other regimes. On CLINC150, a low-margin SCA fallback gives only a narrow marginal gain, while Low BGE top-1 is the most consistently effective tested routing rule that requires no additional router training. The broader lesson is to select among direct retrieval, discriminative training, generative reranking, and API escalation according to available supervision and deployment cost.

Limitations

The study has several limitations. First, always-on SCA is weaker than direct centroid retrieval on aggregate CLINC150 accuracy, and task-trained MPNet baselines are stronger when supervised re-training is acceptable. Selective SCA produces a positive net change only at one narrow validation-selected low-margin operating point; its confidence interval includes zero at the next target and larger targets are non-positive. Second, the frozen BGE reranker comparison covers four zero-shot candidate-text constructions, not task-tuned or otherwise optimized cross-encoders. Its timing and the SCA timings use different throughput batch sizes and should not be read as strict batch-1 latency rankings. Third, live API routing is evaluated only on CLINC150. BANKING77 and HWU64 test SCA transfer, not cascade generalization, and API caches are complete only through a 20% budget. Fourth, the OOS-aware prompt and selected Qwen3.7-Max version affect cascade behavior, and absolute costs may change with API pricing. Fifth, our evaluation is restricted to public benchmarks and one local serving setup; we do not test on production traffic or evaluate robustness under distribution shift. Finally, failure-graph features add no

consistent oracle gain, and learned-router transfer covers only two small backbones under an oracle diagnostic.

Ethics Statement

This work uses public intent-classification benchmarks and no new user data. The routing framework can reduce large-model calls, but may still inherit errors or biases from both models; deployed OOS handling should be monitored because wrong in-scope assignment can trigger inappropriate downstream actions.

Acknowledgments

This work was supported by Sichuan Science and Technology Program under Grant 2025YFNH0007, the National Major Scientific Instruments and Equipments Development Project of National Natural Science Foundation of China under Grant 62427820, and the Ministry of Education Engineering Research Center Guiding Project for Machine Learning and Industrial Intelligence Applications under Grant SCU2024D013.

We thank the reviewers for their constructive feedback and valuable suggestions that helped improve this work.

References

- Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. [Intent detection in the age of LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1559–1570, Miami, Florida, US. Association for Computational Linguistics.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. [Efficient intent detection with dual sentence encoders](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2024. [FrugalGPT: How to use large language models while reducing cost and improving performance](#). *Transactions on Machine Learning Research*.
- Peng Chen and Bang Wang. 2025. [On collaborating small and large models for few-shot intent detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13937–13953, Suzhou, China. Association for Computational Linguistics.
- Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. 2024. [Hybrid LLM: Cost-efficient and quality-aware query routing](#). In *The Twelfth International Conference on Learning Representations*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Dan Hendrycks and Kevin Gimpel. 2017. [A baseline for detecting misclassified and out-of-distribution examples in neural networks](#). In *International Conference on Learning Representations*.
- Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. 2023. [When does confidence-based cascade deferral suffice?](#) In *Advances in Neural Information Processing Systems*, volume 36, pages 9891–9906.
- Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. [An evaluation dataset for intent classification and out-of-scope prediction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.
- Junhua Liu, Tan Yong Keat, Bin Fu, and Kwan Hui Lim. 2024. [LARA: Linguistic-adaptive retrieval-augmentation for multi-turn intent classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1096–1106, Miami, Florida, US. Association for Computational Linguistics.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. [Energy-based out-of-distribution detection](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475.
- Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. [Benchmarking natural language understanding services for building conversational agents](#). In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 165–183, Singapore. Springer Singapore.
- Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. [In-context learning for text classification with many labels](#). In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. [RouteLLM: Learning to route LLMs with preference data](#). In *The Thirteenth International Conference on Learning Representations*.
- Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [MPNet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.
- Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. [Efficient few-shot learning without prompts](#). *arXiv preprint arXiv:2209.11055*.
- Chengyu Wang, Haojie Pan, Yuan Liu, Kehan Chen, Minghui Qiu, Wei Zhou, Jun Huang, Haiqing Chen, Wei Lin, and Deng Cai. 2021. [MeLL: Large-scale extensible user intent classification for dialogue systems with meta lifelong learning](#). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3649–3659, New York, NY, USA. Association for Computing Machinery.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-Pack: Packed resources for general Chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649, New York, NY, USA. Association for Computing Machinery.
- Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. [Few-shot intent classification and slot filling with retrieved examples](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 734–749, Online. Association for Computational Linguistics.
- Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021. [Deep open intent classification with adaptive decision boundary](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.
- Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. [Discriminative nearest neighbor few-shot intent detection by transferring natural language inference](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082. Association for Computational Linguistics.
- Ziji Zhang, Michael Yang, Zhiyu Chen, Yingying Zhuang, Shu-Ting Pi, Qun Liu, Rajashekar Maragoud, Vy Nguyen, and Anurag Beniwal. 2025. [REIC: RAG-enhanced intent classification at scale](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1072–1080, Suzhou (China). Association for Computational Linguistics.
- Yunhua Zhou, Jiawei Hong, and Xipeng Qiu. 2023. [Towards open environment intent prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2226–2240, Toronto, Canada. Association for Computational Linguistics.

A Additional Detailed Tables

Table 7 gives the composition of the submitted learned-router subsets. Table 8 reports the original oracle threshold comparison, and Table 9 gives model-participation ablations. These are oracle diagnostics unless explicitly marked as real API results.

Model	Route	Routed	OOS in routed	OOS recall	Fixes/reg.
Qwen	10%	550	77.27	42.50	306 / 0
Qwen	20%	1100	64.73	71.20	541 / 0
Gemma	10%	550	72.73	40.00	310 / 1
Gemma	20%	1100	57.55	63.30	497 / 2

Table 7: OOS routing behavior of the real-API learned-router cascades on CLINC150 test. OOS values are percentages except routed counts and OOS fix/regression counts.

Model	Score	Budget	Oracle acc.	Hard prec.	OOS recall
Qwen	LR	10%	83.02	68.91	42.50
Qwen	top-1	10%	83.25	71.27	51.80
Qwen	margin	10%	82.49	63.64	35.30
Qwen	entropy	10%	83.64	75.09	44.40
Qwen	OOS	10%	80.24	41.09	29.40
Qwen	LR	20%	89.18	65.27	71.20
Qwen	top-1	20%	89.62	67.45	83.80
Qwen	margin	20%	88.69	62.82	64.90
Qwen	entropy	20%	89.25	65.64	72.50
Qwen	OOS	20%	84.56	42.18	59.10
Gemma	LR	10%	83.47	73.27	40.00
Gemma	top-1	10%	82.45	63.09	51.80
Gemma	margin	10%	82.62	64.73	35.30
Gemma	entropy	10%	83.18	70.36	44.40
Gemma	OOS	10%	80.27	41.27	29.40
Gemma	LR	20%	89.00	64.27	63.30
Gemma	top-1	20%	88.85	63.55	83.80
Gemma	margin	20%	88.53	61.91	64.90
Gemma	entropy	20%	88.55	62.00	72.50
Gemma	OOS	20%	83.84	38.45	59.10

Table 8: Oracle diagnostic results, not real-API cascade measurements. LR is the learned router; scalar rules route low top-1, low margin, high entropy, or high OOS scores on CLINC150 test. Recall is OOS route recall.

B Additional Experimental Details

All four BGE candidate-text constructions passed structural checks for score direction, query-candidate order, candidate mapping, and truncation. Independent full reruns of the two structured variants reproduced all 11,000 predictions and all 55,000 candidate logits exactly.

Router	Train→Deploy	Budget	Oracle acc.	Hard prec.	OOS recall
Pre-gen	Q→Q	10%	83.11	69.82	50.30
Pre-gen	Q→Q	20%	89.42	66.45	77.60
Pre-gen	G→G	10%	83.42	72.73	48.30
Pre-gen	G→G	20%	89.49	66.73	76.00
Pre-gen	Q→G	10%	82.65	65.09	50.30
Pre-gen	Q→G	20%	89.09	64.73	77.60
Pre-gen	G→Q	10%	83.75	76.18	48.30
Pre-gen	G→Q	20%	89.56	67.18	76.00
Src-FG	Q→G	10%	82.96	68.18	43.90
Src-FG	Q→G	20%	88.51	61.82	70.50
Src-FG	G→Q	10%	83.67	75.45	42.00
Src-FG	G→Q	20%	89.35	66.09	67.10

Table 9: Oracle diagnostic results, not real-API cascade measurements. These model-participation ablations use CLINC150 test. *Pre-gen* uses retrieval and complexity features only at deployment, with no failure graph and no deploy-model prediction; its training labels still come from source-model validation errors, which is why source-to-deploy transfer is reported. In transfer rows the deploy model contributes no error data, so the routed set (and OOS recall) is identical to the source-native rows. *Src-FG* is a separate source-only failure-graph ablation, not a pre-generation setting. Q = Qwen2.5-3B-Instruct, G = Gemma-2-2B-it.

Model	Centroid	$k = 3$	$k = 5$	R@3	R@5
Qwen	93.35	88.87	87.61	97.97	98.71
Gemma	93.35	88.19	86.97	97.97	98.71

Model ($k = 5$)	0 ex.	1 ex.	4 ex.
Qwen	85.97	87.61	87.16
Gemma	81.48	86.97	85.71

Table 10: CLINC150 validation sensitivity (percentages). Centroid directly returns BGE top-1 and is not an SCA candidate-size setting. R@ k is candidate recall.

Reranker/candidate text	Acc.	In-scope	OOS F1	ms/q
BGE: raw label	57.51	65.73	33.52	2.56
BGE: humanized	56.07	68.09	3.92	2.76
BGE: structured	56.13	66.93	13.55	3.89
BGE: structured + 1 ex.	59.89	71.29	15.45	5.55
Qwen SCA	76.13	89.16	29.61	130.7
Gemma SCA	76.15	88.24	35.52	128.8

Table 11: Frozen non-generative reranker comparison on CLINC150 test. Times are throughput-amortized references on one V100, exclude shared retrieval, and use batch sizes 64/2/4 for BGE/Qwen/Gemma.

<i>Qwen</i> Val. target	Test call	Acc.	Net	Gain (95% CI)
0%	0.00	80.71	0	[0.00,0.00]
5%	8.62	81.18	+26	[+0.05,+0.89]
10%	17.16	81.07	+20	[-0.18,+0.91]
20%	30.04	79.93	-43	[-1.47,-0.09]
30%	40.53	78.62	-115	[-2.85,-1.31]
50%	58.89	77.24	-191	[-4.29,-2.64]
100%	100.00	76.13	-252	[-5.44,-3.69]
<i>Gemma</i>				
0%	0.00	80.71	0	[0.00,0.00]
5%	8.62	81.13	+23	[+0.04,+0.80]
10%	17.16	81.22	+28	[0.00,+1.00]
20%	30.04	80.45	-14	[-0.91,+0.38]
30%	40.53	79.51	-66	[-1.91,-0.49]
50%	58.89	77.78	-161	[-3.73,-2.15]
100%	100.00	76.15	-251	[-5.42,-3.69]

Table 12: Low-margin selective SCA on CLINC150 test. Validation thresholds are applied unchanged to test. Net is the change in the number of correct predictions relative to 151-centroid top-1; Gain (95% CI) is the test-utterance-level paired-bootstrap accuracy difference in percentage points.

<i>Qwen</i> Budget	Low BGE	Token conf.	Learned
5%	79.24/50.74	78.96/45.83	79.35/50.00
10%	82.89/70.00	81.38/56.49	82.24/63.93
15%	86.00/82.01	84.13/66.96	84.71/73.68
20%	88.07/87.47	86.13/73.74	87.00/80.99
<i>Gemma</i>			
5%	78.73/52.41	79.53/54.58	79.51/55.62
10%	82.11/70.13	82.20/65.54	82.42/67.39
15%	85.24/82.01	84.35/72.17	84.84/75.68
20%	87.33/87.70	85.76/76.34	86.45/80.18

Table 13: Actual cached-API route-rate curve on CLINC150 as accuracy/OOS F1 (percentages). Exact-setting caches are complete through 20%; larger route rates are not imputed.

B.1 SCA Prompt Template

The reported CLINC150 runs omit retrieval scores and truncate each example to 100 characters. Humanization replaces underscores with spaces; for oos, it returns “out of scope: none of the supported intents match the utterance.” For each of five candidates, the bracketed fields below are replaced by the label, its humanization, and one fixed training utterance. This block is the user-message content; each tokenizer adds its model-specific chat template.

You are an intent classification reranker.
Choose exactly ONE intent label from the
candidate list.
Use the candidate meanings and examples only to
distinguish similar intents.
Do not explain. Output only the label string.
Candidate intents:
1. [label]
 meaning: [humanized label]
 examples:
 - [representative utterance]
 ...
User utterance:
[query]
Output label:
