

Out-of-Scope-Aware Tool Routing Agents with Selective Invocation and Calibrated Rejection

Matthew Taylor¹, Min Wu², Megan Harris³

¹Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA

²Department of Computer Science, New York University, New York, NY, USA

³Department of Computer Science, University of California, Davis, CA, USA

m.taylor.tech@outlook.com

Abstract

Tool-using agents need a router that selects the correct capability while refusing requests outside the available catalog. This study formulates routing on CLINC150 as joint 150-way tool classification, out-of-scope (OOS) detection, and selective invocation. The proposed Calibrated Selective Invocation Router (CSIR) combines word and character TF-IDF features, a linear support vector machine route head, temperature-scaled route probabilities, a 151-class OOS specialist, class-prototype distance, and a calibrated rejection gate. The official partitions provide 15,000/3,000/4,500 in-scope training/validation/test utterances and 100/100/1,000 OOS utterances. Disjoint validation subsets calibrate probabilities and select thresholds before frozen evaluation on 5,500 test queries. The hybrid router achieves 92.73% top-1, 97.36% top-3, and 98.11% top-5 accuracy. Temperature scaling reduces expected calibration error from 0.8705 to 0.0212. OOS fusion reaches 0.9585 AUROC, 0.8593 AUPR-OOS, and 0.1682 FPR at 95% OOS recall. Under moderate costs, CSIR accepts 75.11% of queries at 94.55% precision, rejects 90.90% of OOS inputs, and yields 0.5046 utility per query, exceeding 0.4796 for a maximum-probability gate and -0.3287 for always invoking. Paired bootstrap intervals support the routing, OOS, and utility gains, while warmed batch inference requires 0.0529 ms per query.

Keywords— Tool Routing, Autonomous Agents, Intent Classification, Out-Of-Scope Detection, Selective Prediction, Confidence Calibration, Abstention, CLINC150

INTRODUCTION

Tool-using agents expose specialized capabilities through a routing layer. MRKL provides a modular interface between language models and discrete resources [1], while hybrid-cloud deployment research shows why capability placement, latency, and cost must be considered together [2]. ReAct couples reasoning with actions [3], and execution-guided CI diagnosis illustrates the need to ground action selection in observable feedback [4]. Toolformer learns when to call tools [5], whereas retrieval-summary integration for code intelligence separates candidate retrieval from downstream explanation [6]. Gorilla emphasizes accurate API choice [7], and evidence-grounded RAG demonstrates that retrieved support should constrain the answer or action that follows [8].

The broader augmented-language-model literature treats external retrieval and tools as extensions of parametric knowledge [9]. Long-document RAG for contractual and insurance analysis further shows that routing must preserve the relevant evidence span rather than merely identify a broad domain [10]. HuggingGPT coordinates heterogeneous models [11], while retrieved normal prototypes have been used to explain OpenStack failure-injection anomalies [12]. ToolLLM scales training to large API collections [13], and selective refusal in HDFS anomaly analysis connects retrieval quality with an explicit decision not to answer when evidence remains ambiguous

[14]. These systems motivate a router that ranks candidates and controls whether any candidate should execute.

CLINC150 supplies 150 intents and explicit OOS examples for this problem [15]. Language context can affect how technical concepts are understood, as work on bilingual computer-science education documents [16], so a fixed English intent inventory should not be treated as a universal capability boundary. Snips established efficient intent recognition for embedded assistants [17], and e-commerce classification and recommendation pipelines show similar dependence on fine-grained category distinctions [18]. Banking77 provides another dense intent taxonomy [19]. Outside assistants, cancer-image classification [20] and AI-based minimal-residual-disease evaluation [21] illustrate the same operational point: a label is useful only when its uncertainty and downstream consequence are interpreted together.

OOS detection prevents a closed-set router from forcing every query onto a known tool. Maximum softmax probability is a standard confidence baseline [22], and human-uncertainty distillation shows that supervision about ambiguity can improve calibration [23]. Class-conditional distance provides geometric evidence [24], paralleling anomaly detection with deep autoencoders in IoT traffic [25]. Outlier exposure shapes boundaries with auxiliary anomalies [26], while residual fusion has been used to identify noisy-neighbor VM degradation [27].

Energy scores offer a logit-based alternative [28], and predictive DDoS mitigation demonstrates why false acceptance can carry concrete system cost [29]. Open-set recognition consequently treats a strong closed-set representation as necessary but not sufficient for unknown rejection [30].

Robust scope decisions often combine complementary signals. Perception fusion in autonomous driving is one example of multi-source decision making [31], while margin-based unknown-intent detection directly separates unsupported requests [32]. Attention-enhanced image detection shows how representation changes can sharpen difficult class boundaries [33], and adaptive decision boundaries tailor rejection to each known intent [34]. Confidence-calibrated late fusion provides a related uncertainty-aware design [35]. Self-supervised OOS training [36], domain transfer through approximately shared features [37], and energy-based unknown-intent learning [38] all seek robustness when test inputs depart from the training distribution.

A deployed router must also remain stable after release. Multi-signal drift monitoring and safe rollback provide one operational pattern [39]; soft labels for generated OOD examples provide another way to represent uncertain boundaries [40]. AIOps fault analysis emphasizes that detection, explanation, and remediation should remain separable stages [41]. Accordingly, CSIR treats tool ranking, scope estimation, and invocation as distinct outputs that can be monitored and retuned without retraining the entire pipeline.

Calibration turns a score ordering into a probability usable by policy. Platt scaling maps support-vector outputs to probabilities [42]. Spectral rank-aggregation theory clarifies that accurate ordering alone does not establish uncertainty [43], and comparative calibration studies make the same distinction empirically [44]. Uncertainty quantification for spectral estimators further separates point estimation from confidence assessment [45]. Temperature scaling offers a simple multiclass correction [46], while uncertainty-aware macro-market fusion shows how heterogeneous predictive signals can be combined without treating them as equally reliable [47].

Selective prediction converts calibrated evidence into an action. Risk-coverage theory formalizes the reject option [48], and budget-constrained auto-bidding demonstrates why thresholds should reflect asymmetric penalties [49]. SelectiveNet learns coverage-aware rejection [50], while counterfactual learning-to-rank emphasizes evaluation under the policy that actually exposes items or actions [51]. Digital-twin dispatching likewise couples predictive quality to operational decisions and costs [52]. For tool routing, the analogous question is not only which tool ranks first, but whether the expected value of invoking it exceeds the value of deferral.

This study evaluates that joint question by mapping every CLINC150 in-scope intent to a separate tool. CSIR uses a hybrid sparse route head, temperature calibration, an explicit 151-class OOS specialist, prototype distance, and a log-odds invocation score. The evaluation covers top-k routing, calibration, OOS discrimination, risk-coverage, cost-sensitive utility, component ablations, error concentration, efficiency, prevalence shift, and paired

bootstrap intervals. The experiments show that explicit OOS supervision supplies most rejection power, prototype evidence adds a smaller consistent gain, and calibrated fusion improves utility over confidence-only and OOS-only gates without changing the underlying tool ranking.

METHOD

A. Task Formulation and Decision Semantics

Let $T = \{t_1, \dots, t_{150}\}$ be the supported tool catalog, with one tool per in-scope intent. For utterance x , the route head returns scores $f(x)$ in $\mathbb{R}^{150}$ and ranking $\pi(x)$; the top prediction is $\hat{t} = \arg\max_j f_j(x)$. Top-k routing succeeds when the gold tool occurs among the k highest scores. Invocation is a separate action $a(x)$ in $\{\text{invoke}, \text{reject}\}$. A correct outcome requires invoking the gold tool for an in-scope query, whereas wrong-tool invocation and OOS invocation are recorded separately.

Coverage is the fraction of all queries invoked; accepted precision is the fraction of invoked queries routed correctly; in-scope coverage measures retained availability; and OOS rejection measures containment of unsupported requests. Utility assigns u_{correct} , u_{wrong} , u_{OOS} , and u_{reject} to the four outcomes. For safe score $s(x)$, a threshold τ selected on validation data invokes when $s(x) \geq \tau$ and is then frozen for test evaluation.

B. Dataset and Experimental Protocol

Experiments use the official CLINC150 partitions [15]. Training contains 15,000 in-scope utterances (100 per intent), validation 3,000 (20 per intent), and test 4,500 (30 per intent). The OOS partitions contain 100 training, 100 validation, and 1,000 test utterances. Table I reports composition and token length. Each intent is treated as one callable tool, so no labels or examples are removed.

TABLE I Dataset composition and utterance length

Split	Examples	Intent classes	OOS examples	Mean tokens	Median	Min	Max
train	15000	150	0	8.34	8	1	28
val	3000	150	0	8.32	8	1	24
test	4500	150	0	8.19	8	1	25
oos_train	100	0	100	7.06	7	3	13
oos_val	100	0	100	7.16	7	1	22
oos_test	1000	0	1000	8.75	8	1	20

Within every intent, the first ten validation utterances form a 1,500-example calibration set and the remaining ten form a 1,500-example gate set. OOS training supports the OOS specialist and scalar calibrator; OOS validation joins the gate set for fusion and threshold selection. Frozen evaluation uses 4,500 in-scope plus 1,000 OOS test queries (18.18% OOS). Figure 2 summarizes the protocol. Seed 20260822 controls prevalence resampling and bootstrap tests.

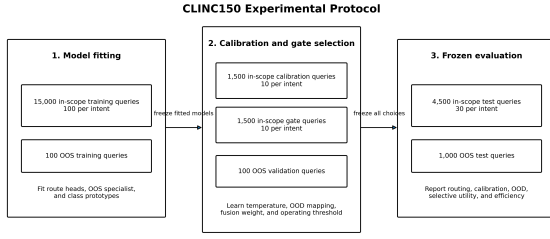


Fig. 2. Experimental protocol. Calibration, gate selection, and frozen testing use disjoint in-scope validation subsets and the official OOS partitions.

C. Text Representation and Closed-Set Route Head

CSIR uses two TF-IDF views, following term-weighting principles that emphasize discriminative terms [53]. The word view uses lowercase unigrams and bigrams, accent stripping, sublinear term frequency, $\min_df=2$, $\max_df=0.995$, and L2 normalization. The character-within-word view uses 3-5 grams, $\min_df=2$, sublinear frequency, and L2 normalization. The fitted vocabularies contain 11,860 and 20,521 features.

The sparse views are concatenated and normalized to 32,381 dimensions. A linear support vector machine learns 150 one-versus-rest margins, using $C=2.0$, tolerance 10^{-4} , automatic primal-dual selection, and 8,000 maximum iterations. The margin objective follows support-vector classification [54], and the sparse implementation follows LIBLINEAR [55].

Baselines are word TF-IDF with multinomial naive Bayes ($\alpha=0.08$), word TF-IDF with linear SVM, character TF-IDF with linear SVM, and the hybrid word-character SVM. All settings are fixed across experiments and listed in Table II; route rankings are descending SVM decision scores.

TABLE II Model and training configuration

Component	Configuration
Word representation	TF-IDF; word 1-2 grams; $\min_df=2$; $\max_df=0.995$; sublinear TF; L2; 11,860 features
Character representation	TF-IDF char_wb 3-5 grams; $\min_df=2$; sublinear TF; L2; 20,521 features
Hybrid representation	Sparse concatenation followed by L2 normalization; 32,381 features
Naive Bayes baseline	MultinomialNB; $\alpha=0.08$
Route SVMs	LinearSVC; $C=2.0$; $\text{tol}=1e-4$; $\text{dual}=auto$; $\text{max_iter}=8,000$
OOS specialist	151-class LinearSVC with the same SVM settings
Route calibration	Temperature scaling on 1,500 in-scope validation utterances
OOS fusion	Validation-selected $\alpha=0.85$ explicit margin + 0.15 prototype distance
OOS probability	Balanced one-dimensional logistic regression; $C=1.0$; $\text{max_iter}=3,000$
Random seed	20260822

D. Route-Probability Calibration

SVM margins are rankings rather than probabilities. CSIR fits one positive temperature T on the 1,500 calibration utterances by minimizing multiclass negative log-likelihood, yielding $p_j(x)$ proportional to $\exp(f_j(x)/T)$. This post-hoc transformation follows temperature scaling [46] and the broader calibration tradition of sigmoid

mapping and empirical reliability analysis [42], [44]. Test calibration is measured with 15-bin expected calibration error (ECE), negative log-likelihood, multiclass Brier score, and mean top-1 confidence.

E. Out-of-Scope Evidence and Probability Mapping

The explicit specialist is a 151-class linear SVM trained on 15,000 in-scope and 100 OOS training utterances. Its OOS margin is $m(x) = g_OOS(x) - \max_j g_j(x)$. The route head remains a 150-tool model, so OOS supervision cannot alter the reported closed-set ranking.

The geometric score averages normalized hybrid vectors into one centroid per tool and computes $d(x) = -\max_j \cos(h(x), c_j)$. Larger values indicate weak similarity to every known tool, following the class-conditional distance principle [24]. This sparse centroid construction avoids covariance estimation in 32,381 dimensions.

Each signal is standardized on the gate mixture. Validation AUROC selects $z_OOD(x) = 0.85 z_m(x) + 0.15 z_d(x)$, and balanced one-dimensional logistic regression maps the fused score to $p_OOD(x)$. Comparisons include 1 - maximum route probability [22], entropy, prototype distance, raw energy [28], and the explicit margin. No additional OOS utterances are synthesized.

$$z_OOD(x) = 0.85 z_m(x) + 0.15 z_d(x) \quad (1)$$

F. Calibrated Selective Invocation

The safe-invocation score is $s(x) = \text{logit}(p_route(x)) - \text{logit}(p_OOD(x))$, where $p_route(x)$ is the largest calibrated tool probability. High values require both confident identity and low scope risk. Log-odds subtraction preserves more dynamic range than multiplying two probabilities. Figure 1 shows the complete architecture.

$$s(x) = \text{logit}(p_route(x)) - \text{logit}(p_OOD(x)) \quad (2)$$

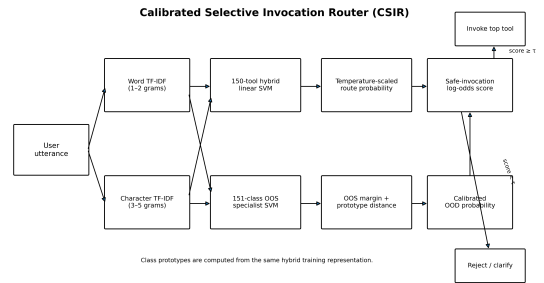


Fig. 1. CSIR architecture. The route branch ranks 150 tools and produces a temperature-scaled confidence; the OOS branch combines an explicit 151-class margin with prototype distance. Calibrated log-odds determine invocation or rejection.

Validation selects either target-precision thresholds (90%, 95%, 97%, 99%) or utility-maximizing thresholds for Table III. The primary moderate regime assigns +1 to a correct invocation, -3 to a wrong known tool, -5 to an OOS invocation, and -0.2 to rejection. This cost-aware reject option follows selective prediction [48], [50] and is analogous to risk-constrained decision policies in auto-bidding [49] and digital-twin dispatching [52]. Comparisons hold route predictions fixed while changing only the gate.

TABLE III Utility regimes used for threshold selection

Regime	Correct	Wrong in-scope	OOS invocation	Reject
Low consequence	+1	-1	-2	-0.1
Moderate consequence	+1	-3	-5	-0.2
High consequence	+1	-6	-10	-0.3
Expensive abstention	+1	-3	-5	-0.8

G. Evaluation Metrics, Statistical Tests, and Efficiency

Routing reports top-1 accuracy, macro F1, and top-3/top-5 accuracy. OOS evaluation reports AUROC, AUPR-OOS, FPR@95TPR, validation-selected OOS F1, balanced accuracy, OOS recall, and in-scope false rejection. Selective evaluation reports coverage, accepted precision, in-scope coverage, OOS rejection, rejection of wrong in-scope predictions, risk-coverage curves, and utility. Prevalence analysis applies the fixed moderate threshold to 20 resampled 5,000-query mixtures at each OOS rate from 1% to 60%.

Paired bootstrap tests compare hybrid versus character-SVM top-1 accuracy (1,000 resamples), CSIR versus maximum-probability OOS AUROC (300), and moderate utility for the same two gates (1,000). Efficiency is the median of five warmed full-test batches including vectorization. Serialized size includes both vectorizers, the route and OOS heads, prototypes, calibrators, fusion weight, and temperature.

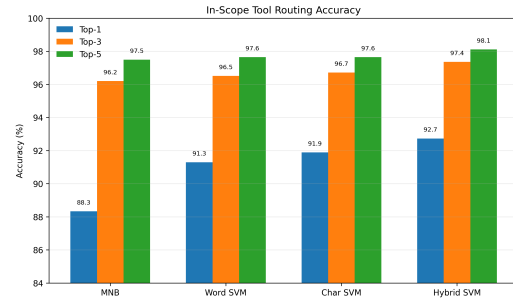
RESULTS AND DISCUSSION

A. In-Scope Tool Routing

Table IV and Figure 3 show top-1 accuracy of 88.33% for multinomial naive Bayes, 91.29% for word SVM, 91.89% for character SVM, and 92.73% for the hybrid route head, whose macro F1 is 92.64%. Hybrid top-3 and top-5 accuracy reach 97.36% and 98.11%. The 0.844-point gain over character SVM has paired 95% interval [0.333, 1.378] and $p=0.002$. Character features preserve lexical variants, while word bigrams distinguish action-bearing phrases.

TABLE IV Closed-set tool-routing performance on 4,500 in-scope test utterances

Method	Top-1 (%)	Macro F1 (%)	Top-3 (%)	Top-5 (%)
Word TF-IDF + MNB	88.33	88.16	96.20	97.49
Word SVM	91.29	91.19	96.51	97.64
Char SVM	91.89	91.78	96.71	97.64
Hybrid word-char SVM (CSIR route head)	92.73	92.64	97.36	98.11

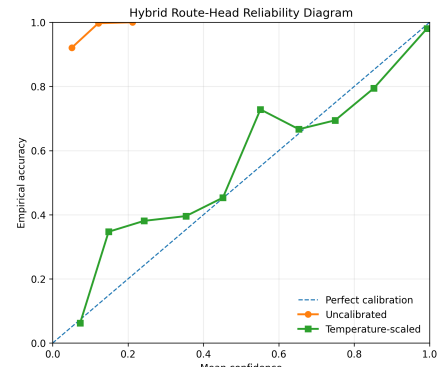
**Fig. 3.** Top-k in-scope tool-routing accuracy. The hybrid route head gives the strongest top-1, top-3, and top-5 performance.

B. Probability Calibration

Table V and Figure 5 show that unit-temperature softmax is severely underconfident: hybrid mean confidence is 0.0568 despite 0.9273 accuracy, producing ECE 0.8705, NLL 3.0450, and Brier 0.8983. Temperature $T=0.1609$ raises mean confidence to 0.9340 and reduces ECE to 0.0212, NLL to 0.3412, and Brier to 0.1116. Rankings are unchanged, but the top probability becomes usable by the invocation policy.

TABLE V Probability calibration of SVM route heads

Method	State	T	ECE	NLL	Brier	Mean confidence
Word SVM	uncalibrated	1.0000	0.8606	3.1455	0.9071	0.0522
Word SVM	temperature-scaled	0.1616	0.0259	0.4176	0.1335	0.9202
Char SVM	uncalibrated	1.0000	0.8495	2.8814	0.8763	0.0694
Char SVM	temperature-scaled	0.1809	0.0259	0.3841	0.1263	0.9216
Hybrid SVM	uncalibrated	1.0000	0.8705	3.0450	0.8983	0.0568
Hybrid SVM	temperature-scaled	0.1609	0.0212	0.3412	0.1116	0.9340

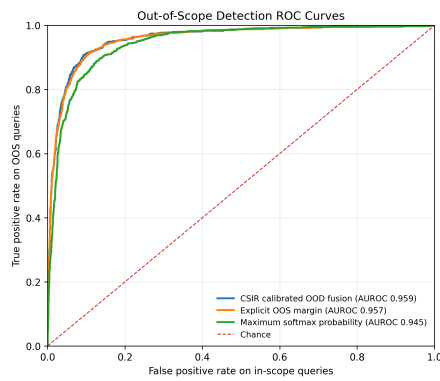
**Fig. 5.** Reliability of the hybrid route head before and after temperature scaling. Raw softmax values are strongly underconfident; calibrated values track empirical accuracy.

C. Out-of-Scope Detection

Table VI and Figure 4 compare six OOS scores. Maximum probability reaches 0.9450 AUROC, 0.7963 AUPR-OOS, and 0.2311 FPR@95TPR; entropy reaches 0.9488 AUROC, and prototype distance 0.9032. Raw energy is weak (0.5810 AUROC) because the route SVM was not trained to shape energy magnitudes.

TABLE VI Out-of-scope detection on the 5,500-query mixed test set

Method	AUROC	AUPR-OOS	FPR@95	OOS F1	Bal. acc.	OOS recall	ID false reject
Maximum softmax probability	0.9450	0.7963	0.2311	0.7159	0.8021	0.6350	0.0309
Predictive entropy	0.9488	0.8127	0.2276	0.7131	0.7965	0.6190	0.0260
Prototype distance	0.9032	0.6536	0.4004	0.5352	0.6938	0.4220	0.0344
Energy score	0.5810	0.1929	0.7618	0.3594	0.6050	0.8560	0.6460
Explicit OOS margin	0.9570	0.8546	0.1762	0.7378	0.8094	0.6400	0.0211
CSIR calibrated OOD fusion	0.9585	0.8593	0.1682	0.7147	0.7927	0.6050	0.0196

**Fig. 4.** OOS receiver operating characteristic curves. CSIR fusion improves on explicit-margin and maximum-probability scores.

The explicit 151-class margin is the strongest single signal at 0.9570 AUROC and 0.1762 FPR@95TPR. Adding 15% prototype evidence yields 0.9585 AUROC, 0.8593 AUPR-OOS, and 0.1682 FPR@95TPR. Relative to maximum probability, the AUROC gain is 0.0136 with interval [0.00999, 0.01690], and FPR@95TPR falls 27.2%.

D. Selective Invocation and Risk-Coverage Behavior

Table VII transfers validation thresholds unchanged. Always invoking has 100% coverage but only 75.87% precision because it executes all 1,000 OOS queries and 327 wrong in-scope routes.

TABLE VII CSIR selective-invocation operating points

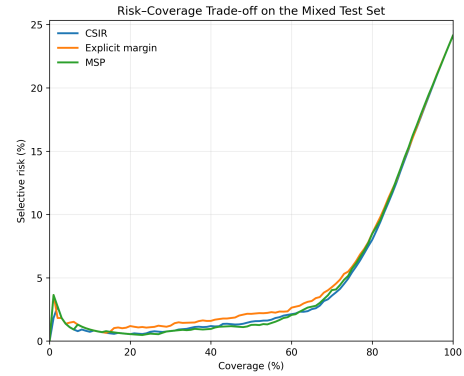
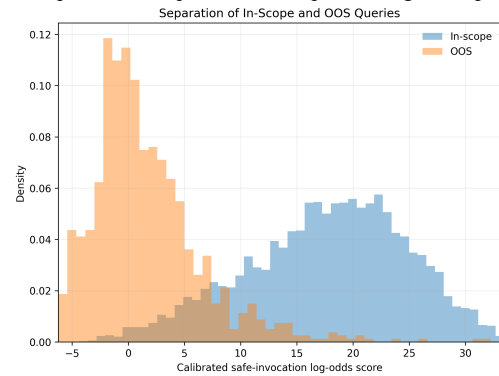
Policy	Threshold	Coverage (%)	Accepted precision (%)	ID coverage (%)	OOS rejection (%)	Wrong-ID rejection (%)
always invoke	-inf	100.00	75.87	100.00	0.00	0.00
target precision 90%	-0.295	91.80	82.51	99.27	41.80	7.95
target precision 95%	5.976	78.60	92.76	92.96	86.00	47.09
target precision 97%	8.760	72.60	95.64	87.11	92.70	69.11
target precision 99%	15.532	53.20	98.39	64.62	98.20	91.13

Policy	Threshold	Coverage (%)	Accepted precision (%)	ID coverage (%)	OOS rejection (%)	Wrong-ID rejection (%)
cost-optimal moderate	7.646	75.11	94.55	89.78	90.90	59.02

The 95%, 97%, and 99% target policies accept 78.60%, 72.60%, and 53.20% of test queries at 92.76%, 95.64%, and 98.39% precision, with OOS rejection of 86.0%, 92.7%, and 98.2%. The 90% target transfers less reliably because its threshold lies in the low-score tail represented by only 100 OOS validation examples.

The primary moderate policy accepts 4,131 queries: 75.11% coverage, 94.55% precision, 89.78% in-scope coverage, and 90.90% OOS rejection. It also rejects 59.02% of wrong in-scope predictions.

Figure 6 shows lower middle-coverage risk for CSIR than maximum-probability and explicit-margin gates; Figure 7 shows why, with OOS scores concentrated below the in-scope distribution. The remaining overlap contains ambiguous known requests and unsupported variants close to broad tools.

**Fig. 6.** Selective risk-coverage trade-off on the mixed test set. Risk counts accepted OOS requests and accepted wrong in-scope routes.**Fig. 7.** Distribution of calibrated safe-invocation scores for in-scope and OOS test queries. Higher scores favor invocation.

E. Cost-Sensitive Utility

Table VIII reports validation-tuned utility. CSIR leads under low, moderate, and high consequences with 0.6169, 0.5046, and 0.3947 per query. In the moderate regime it exceeds maximum-probability gating by 0.02495 (95% interval [0.01054, 0.04081], $p=0.002$) and always-invoke by 0.8333. With rejection cost -0.8, maximum probability leads narrowly, 0.3160 versus 0.3099. This dependence on consequences resembles credit-risk decision rules [56] and

motivates interfaces that expose both risk and uncertainty rather than a single categorical output [57].

TABLE VIII *Utility per query under validation-tuned policies*

Method	Low	Moderate	High	Expensive abstention
Always invoke	0.3356	-0.3287	1.4162	-0.3287
MSP gate	0.6001	0.4796	0.3536	0.3160
Prototype gate	0.4738	0.2573	0.1278	-0.0101
Explicit OOS-margin gate	0.6080	0.4943	0.3518	0.3046
CSIR calibrated selective gate	0.6169	0.5046	0.3947	0.3099

F. Component Ablations

Table IX shows that explicit OOS supervision raises AUROC from 0.9450 to 0.9570; the 15% prototype component adds 0.00150 AUROC and lowers FPR@95TPR by 0.0080. The gain is small but consistent, so geometry complements rather than replaces the explicit margin.

TABLE IX *Ablation of out-of-scope evidence*

Variant	AUROC	AUPR-OOS	FPR@95	OOS F1	OOS recall	ID false reject
MSP only	0.9450	0.7963	0.2311	0.7159	0.6350	0.0309
Prototype only	0.9032	0.6536	0.4004	0.5352	0.4220	0.0344
Explicit margin only	0.9570	0.8546	0.1762	0.7378	0.6400	0.0211
Explicit/prototype fusion (0.85/0.15)	0.9585	0.8593	0.1682	0.7147	0.6050	0.0196

Table X compares invocation formulas under moderate costs. Route confidence, OOS probability, multiplication, and minimum confidence yield utilities 0.4796, 0.4762, 0.4927, and 0.4864; log-odds fusion reaches 0.5046. Such policy-level comparison is important in data-driven marketing decisions [58], and a post-routing explanation card can present the selected tool, alternatives, and rejection evidence without changing the gate [59].

TABLE X *Invocation-score ablation under the moderate cost regime*

Variant	Coverage (%)	Accepted precision (%)	ID coverage (%)	OOS rejection (%)	Wrong-ID rejection (%)	Utility
Route confidence only	75.09	94.26	88.78	86.50	68.81	0.4796
OOD probability only	76.84	93.33	91.42	88.80	48.01	0.4762
Multiplicative confidence	74.73	94.62	88.69	88.10	68.81	0.4927
Minimum confidence	74.93	94.44	88.76	87.30	68.81	0.4864
CSIR log-odds difference	75.11	94.55	89.78	90.90	59.02	0.5046

G. Error Analysis

Table XI lists the twelve most frequent directed confusions. `redeem_rewards` -> `rewards_balance` and `calendar` ->

`calendar_update` each occur eight times; `improve_credit_score` -> `credit_score` occurs seven. Most errors contrast action with status or a broad intent with a specific operation.

TABLE XI *Most frequent directed in-scope routing confusions*

True tool	Predicted tool	Count
<code>redeem_rewards</code>	<code>rewards_balance</code>	8
<code>calendar</code>	<code>calendar_update</code>	8
<code>improve_credit_score</code>	<code>credit_score</code>	7
<code>todo_list_update</code>	<code>todo_list</code>	5
<code>change_user_name</code>	<code>user_name</code>	5
<code>food_last</code>	<code>expiration_date</code>	5
<code>pin_change</code>	<code>change_ai_name</code>	5
<code>replacement_card_duration</code>	<code>new_card</code>	4
<code>transactions</code>	<code>spending_history</code>	4
<code>distance</code>	<code>directions</code>	4
<code>shopping_list</code>	<code>todo_list</code>	4
<code>pto_request</code>	<code>pto_request_status</code>	4

The pattern explains the top-k gain: neighboring labels share vocabulary, so the correct tool often remains in the candidate set. A schema check or clarification question can resolve these pairs more directly than a global threshold. Table XII shows that text receives seven OOS false accepts and directions six; broad tools attract the remaining errors. This concentration matters in financial-service analytics, where apparently familiar language may still fall outside an authorized action boundary [60].

TABLE XII *Tool destinations for OOS false acceptances at the moderate threshold*

Predicted tool	False accepts
<code>text</code>	7
<code>directions</code>	6
<code>who_made_you</code>	3
<code>pto_used</code>	3
<code>direct_deposit</code>	3
<code>definition</code>	3
<code>recipe</code>	3
<code>distance</code>	3
<code>order</code>	3
<code>make_call</code>	3
<code>food_last</code>	2
<code>sync_device</code>	2

Table XIII gives accepted examples such as "ignore call" and "deny incoming phone call," which resemble known communication tools but request unsupported operations. Required-argument and authorization checks should therefore follow routing and precede execution.

TABLE XIII *Representative accepted errors at the moderate threshold*

Type	Utterance	True	Predicted	Safe score
OOS false acceptance	ignore call	OOS	<code>make_call</code>	33.849

Type	Utterance	True	Predicted	Safe score
OOS false acceptance	give me the weather forecast for today	OOS	weather	31.564
OOS false acceptance	read text	OOS	text	30.909
OOS false acceptance	deny incoming phone call	OOS	make_call	26.009
OOS false acceptance	read my friend's text message	OOS	text	23.880
in-scope wrong acceptance	repeat what the weather will be like	transfer	weather	18.331
in-scope wrong acceptance	how is a vacation request done	pto_request	pto_request_status	8.385
in-scope wrong acceptance	can i purchase a vacation using this	pto_request	order	8.488
in-scope wrong acceptance	vacation request information	pto_request	pto_request_status	16.348
in-scope wrong acceptance	vacation request please	pto_request	pto_request_status	13.200

H. Efficiency and OOS-Prevalence Sensitivity

Table XIV reports that the word SVM fits in 3.27 s at 0.0086 ms/query, the hybrid head in 3.67 s at 0.0369 ms/query, and complete CSIR in 7.34 s at 0.0529 ms/query with 78.08 MB uncompressed size. The 0.0159-ms increment covers the OOS classifier, centroid similarity, scalar calibration, and safe-score calculation. Compact proxy models for edge quality prediction pursue a similar latency-versus-fidelity trade-off [61].

TABLE XIV Training time, end-to-end batch latency, and serialized size

Method	Fit (s)	ms/query	Size (MB)	Features
Word TF-IDF + MNB	0.171	0.0086	27.59	11860
Word SVM	3.266	0.0080	14.02	11860
Char SVM	2.666	0.0285	24.16	20521
Hybrid SVM	3.669	0.0369	38.17	32381
CSIR complete router	7.340	0.0529	78.08	32381

Measurements include vectorization but exclude network transport and downstream APIs. Resource budgets remain deployment-specific, as constrained data-driven budgeting frameworks also emphasize [62].

Table XV and Figure 8 apply the fixed moderate threshold as OOS prevalence changes. At 1% OOS, utility is 0.7406 with 96.48% precision; at 20%, utility is 0.4788; at 60%, utility is -0.0793 although OOS rejection remains 90.88%. The score is stable, but residual false accepts dominate high-OOS mixtures. Progressive prioritization systems for document rendering similarly show that a fixed ordering policy must be reevaluated when workload composition changes [63].

TABLE XV Sensitivity to OOS prevalence using the fixed moderate CSIR threshold

OOS prevalence (%)	Utility mean	Utility SD	Coverage (%)	Accepted precision (%)	OOS rejection (%)
1	0.7406	0.0094	88.99	96.48	90.40
5	0.6862	0.0109	85.69	96.14	90.28
10	0.6179	0.0041	81.73	95.59	90.72

OOS prevalence (%)	Utility mean	Utility SD	Coverage (%)	Accepted precision (%)	OOS rejection (%)
20	0.4788	0.0031	73.65	94.28	90.90
40	0.2046	0.0134	57.41	90.73	91.07
60	-0.0793	0.0148	41.41	83.90	90.88

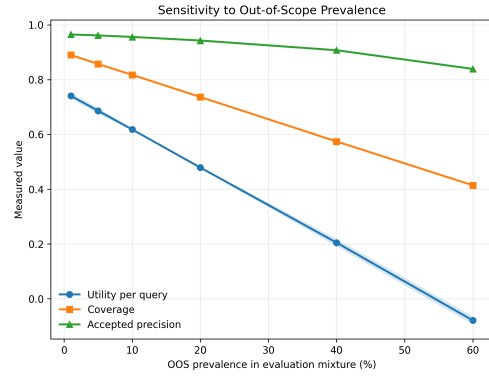


Fig. 8. Sensitivity of fixed-threshold utility, coverage, and accepted precision to OOS prevalence. OOS rejection remains near 91% across mixtures.

I. Statistical Summary and Practical Interpretation

Table XVI consolidates paired bootstrap results. Hybrid routing improves top-1 accuracy over character SVM by 0.00844 with interval [0.00333, 0.01378]; fusion improves OOS AUROC over maximum probability by 0.01356 with interval [0.00999, 0.01690]; and CSIR improves moderate utility by 0.02495 with interval [0.01054, 0.04081]. Every interval excludes zero.

TABLE XVI Paired bootstrap comparisons

Comparison	Estimate	95% CI low	95% CI high	p
Hybrid route head vs Char SVM top-1 accuracy	0.008444	0.003328	0.013778	0.002
CSIR OOD fusion vs MSP AUROC	0.013560	0.009987	0.016901	<0.0034
CSIR vs MSP moderate utility per query	0.024945	0.010544	0.040807	0.002

The evidence supports a three-layer design: high-recall ranking, explicit plus geometric scope evidence, and a calibrated cost-aligned gate. CSIR precedes rather than replaces schema validation, authorization, planning, and execution.

IV. LIMITATIONS

The evaluation uses one English intent benchmark with short utterances. It does not establish performance for multilingual dialogue, dynamically added tools, schema-grounded routing, or structured arguments. The bilingual education evidence cited earlier motivates this concern but does not substitute for a multilingual routing test.

Sparse linear models provide fast deterministic inference but are not compared with semantic encoders or generative routers. OOS supervision contains only 100 training and 100 validation examples, and prevalence experiments change class mixture rather than unsupported-language composition. Utility values are analytical, latency excludes networking and tool execution, and bootstrap intervals cover sampling variation within fixed partitions rather than

benchmark choice, annotation uncertainty, or deployment drift.

CONCLUSION

CSIR treats 150-tool routing as ranking, calibration, OOS detection, and selective action. The hybrid head achieves 92.73% top-1 and 98.11% top-5 accuracy; calibrated ECE is 0.0212; OOS fusion reaches 0.9585 AUROC; and the moderate policy accepts 75.11% of queries at 94.55% precision while rejecting 90.90% of OOS requests. The central result is that reliable tool use requires an explicit decision not to invoke: candidate recall identifies plausible capabilities, whereas calibrated scope evidence and cost-aware thresholds determine whether execution is justified.

REFERENCES

- [1] E. Karpas et al., "MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning," arXiv:2205.00445, 2022.
- [2] Q. Xin, "Hybrid cloud architecture for efficient and cost-effective large language model deployment," *J. Inf. Syst. Informatics*, vol. 7, no. 3, pp. 2182-2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.
- [3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. R. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *Proc. ICLR*, 2023.
- [4] H. Zhang, "LLM-driven CI failure diagnosis and automated repair: From GitHub Actions logs to patch recommendation," *J. Technol. Informatics Eng.*, vol. 4, no. 1, pp. 190-214, Apr. 2025, doi: 10.51903/jtie.v4i1.484.
- [5] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," in *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 68539-68551, 2023.
- [6] Y. Li, "Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset," *J. Adv. Comput. Syst.*, vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.
- [7] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with massive APIs," arXiv:2305.15334, 2023.
- [8] S. Zhou, Z. Li, and E. Wang, "Evidence-grounded RAG for tokenized trade receivable disclosure QA under U.S. capital market standards," *J. Adv. Comput. Syst.*, vol. 3, no. 7, pp. 41-57, Jul. 2023, doi: 10.69987/JACS.2023.30704.
- [9] G. Mialon et al., "Augmented language models: A survey," *Trans. Mach. Learn. Res.*, 2023.
- [10] S. Zhou, Z. Li, and E. Wang, "Long-document RAG for contractual and insurance clause analysis in receivables RWA structures," *J. Adv. Comput. Syst.*, vol. 4, no. 8, pp. 88-104, Aug. 2024, doi: 10.69987/JACS.2024.40810.
- [11] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face," in *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 38154-38180, 2023.
- [12] Q. Xin, "Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes," *Emerg. Inf. Sci. Technol.*, vol. 6, no. 2, pp. 125-146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.
- [13] Y. Qin et al., "ToolLLM: Facilitating large language models to master 16000+ real-world APIs," arXiv:2307.16789, 2023.
- [14] J. Nie and D. Zheng, "Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal," *J. Adv. Comput. Syst.*, vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.
- [15] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proc. EMNLP-IJCNLP*, 2019, pp. 1311-1316.
- [16] G. S. Raj et al., "Impact of bilingual CS education on student learning and engagement in a data structures course," in *Proc. 19th Koli Calling Int. Conf. Comput. Educ. Res.*, 2019, pp. 1-10, doi: 10.1145/3364510.3364518.
- [17] Coucke et al., "Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces," arXiv:1805.10190, 2018.
- [18] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," *Appl. Comput. Eng.*, vol. 64, no. 1, pp. 143-149, May 2024, doi: 10.54254/2755-2721/64/20241365.
- [19] Casanueva, T. Temcinan, D. Gerz, M. Henderson, and I. Vulic, "Efficient intent detection with dual sentence encoders," in *Proc. 2nd Workshop NLP for ConvAI*, 2020, pp. 38-45.
- [20] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," *J. Phys.: Conf. Ser.*, vol. 1651, no. 1, Art. no. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.
- [21] Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD evaluation and prediction model for multiple myeloma," *Front. Comput. Intell. Syst.*, vol. 6, no. 3, pp. 127-131, 2024, doi: 10.54097/zj4MnbWW.
- [22] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. ICLR*, 2017.
- [23] Z. S. Zhong, R. Ma, and H. Zhao, "Human-uncertainty distillation for calibrated vision models on CIFAR-10H," *J. Adv. Comput. Syst.*, vol. 3, no. 2, pp. 77-89, Feb. 2023, doi: 10.69987/JACS.2023.30206.
- [24] Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [25] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," *Appl. Comput. Eng.*, vol. 69, pp. 64-70, Jul. 2024, doi: 10.54254/2755-2721/69/20241511.
- [26] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. ICLR*, 2019.
- [27] J. Nie and D. Zheng, "Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion," *J. Adv. Comput. Syst.*, vol. 4, no. 4, pp. 112-123, Apr. 2024, doi: 10.69987/JACS.2024.40409.
- [28] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21464-21475, 2020.
- [29] Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," *Appl. Comput. Eng.*, vol. 64, no. 1, pp. 94-99, May 2024, doi: 10.54254/2755-2721/64/20241350.
- [30] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?" in *Proc. ICLR*, 2022.
- [31] Y. Wang, S. Du, Q. Xin, Y. He, and W. Qian, "Autonomous driving system driven by artificial intelligence perception fusion," *Acad. J. Sci. Technol.*, vol. 9, no. 2, pp. 193-198, Feb. 2024, doi: 10.54097/e0b9ak47.
- [32] T.-E. Lin and H. Xu, "Deep unknown intent detection with margin loss," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5491-5496.
- [33] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACnv and triplet attention," in *Proc. 2nd Int. Conf. Software Eng. Mach. Learn. (SEML)*, 2024; arXiv:2407.09530.

- [34] H. Zhang, H. Xu, and T.-E. Lin, "Deep open intent classification with adaptive decision boundary," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 16, pp. 14374-14382, 2021.
- [35] Q. Xin, "Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning)," *J. Technol. Informatics Eng.*, vol. 4, no. 1, pp. 215-238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.
- [36] L.-M. Zhan, H. Liang, B. Liu, L. Fan, X.-M. Wu, and A. Y. S. Lam, "Out-of-scope intent detection with self-supervision and discriminative training," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 3521-3532.
- [37] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in *Proc. 28th Int. Conf. Artif. Intell. Statist. (AISTATS)*, PMLR, vol. 258, pp. 559-567, 2025.
- [38] Y. Ouyang, J. Ye, Y. Chen, X. Dai, S. Huang, and J. Chen, "Energy-based unknown intent detection with data manipulation," in *Findings ACL-IJCNLP*, 2021, pp. 2852-2861.
- [39] H. Zhang, "DriftGuard: Multi-signal drift early warning and safe re-training/rollback for CTR/CVR models," *J. Adv. Comput. Syst.*, vol. 3, no. 7, pp. 24-40, Jul. 2023, doi: 10.69987/JACS.2023.30703.
- [40] H. Lang, Y. Zheng, J. Sun, F. Huang, L. Si, and Y. Li, "Estimating soft labels for out-of-domain intent detection," in *Proc. EMNLP*, 2022, pp. 261-276.
- [41] Y. He, Y. Pan, Y. Wang, S. Du, and Q. Xin, "Intelligent fault analysis with AIOps technology," *J. Theory Pract. Eng. Sci.*, vol. 4, no. 1, pp. 94-100, Feb. 2024, doi: 10.53469/jtpes.2024.04(01).13.
- [42] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61-74.
- [43] Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," *Inf. Inference: J. IMA*, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaia/iaae020.
- [44] Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 625-632.
- [45] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," *arXiv:2408.05944*, Aug. 2024.
- [46] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1321-1330.
- [47] H. Zhou and K. Zhang, "News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015-2024 FRED panel," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 487-501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.
- [48] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *J. Mach. Learn. Res.*, vol. 11, pp. 1605-1641, 2010.
- [49] H. Zhang, "Risk-aware budget-constrained auto-bidding under first-price RTB: A distributional constrained deep reinforcement learning framework," *J. Adv. Comput. Syst.*, vol. 4, no. 6, pp. 30-47, Jun. 2024, doi: 10.69987/JACS.2024.40603.
- [50] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2151-2159.
- [51] H. Zhang, "Counterfactual learning-to-rank for ads: Off-policy evaluation on the Open Bandit Dataset," *J. Adv. Comput. Syst.*, vol. 5, no. 12, pp. 1-11, Dec. 2025, doi: 10.69987/JACS.2025.51201.
- [52] Zhang, R. Ma, and P. Greg, "Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 337-363, Aug. 2025, doi: 10.51903/jtie.v4i2.501.
- [53] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513-523, 1988.
- [54] Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995.
- [55] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871-1874, 2008.
- [56] Q. Xin, R. Song, Z. Wang, Z. Xu, and F. Zhao, "Enhancing bank credit risk management using the C5.0 decision tree algorithm," *J. Comput. Technol. Appl. Math.*, vol. 1, no. 4, pp. 100-107, Nov. 2024, doi: 10.5281/zenodo.14032041.
- [57] Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," *Int. J. Graphic Des.*, vol. 3, no. 1, pp. 196-210, May 2025, doi: 10.51903/ijgd.v3i1.3715.
- [58] J. Wang, Q. Xin, Y. Liu, J. Wang, and T. Yang, "Predicting enterprise marketing decision making with intelligent data-driven approaches," *J. Ind. Eng. Appl. Sci.*, vol. 2, no. 3, pp. 12-19, Jun. 2024, doi: 10.5281/zenodo.11357252.
- [59] R. Ma, L. Zhang, and T. Song, "Computer-vision-informed visual explanation cards for autonomous-driving traffic-sign alerts: Localization, classification, and retrieved evidence on GTSDb," *Int. J. Graphic Des.*, vol. 3, no. 2, pp. 455-473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.
- [60] T. Yang, Q. Xin, X. Zhan, S. Zhuang, and H. Li, "Enhancing financial services through big data and AI-driven customer insights and risk analysis," *J. Knowl. Learn. Sci. Technol.*, vol. 3, no. 3, pp. 53-62, Jul. 2024, doi: 10.60087/jklst.vol3.n3.p53-62.
- [61] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 447-463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.
- [62] Y. Lu, H. Zhou, and Y. Zhang, "A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling," *J. Technol. Informatics Eng.*, vol. 4, no. 3, pp. 493-520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.
- [63] H. Wang, Y. Ren, and X. Chang, "Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 425-446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.