Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418- 440
DOI: https://doi.org/10.51903/jtie.v5i1.563

IntentRouter-QAC: Lightweight Lexical Routing for Context-Aware Query Autocomplete and Product Retrieval

Haowei Tu[1], Yinchen Shi*[2], Sophia Chen[3]
Email: ogyinchen@gmail.com*
[1]*Information Systems, New York University, NY, USA*
[2]*Computer Science, New York University, NY, USA*
[3]*Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA*
*Corresponding Author

Abstract
Search boxes increasingly function as routing interfaces: a partial query may require a completion, immediate submission, a context-sensitive suggestion, or a fallback action. This paper evaluates IntentRouter-QAC as a lightweight lexical routing framework and tests product retrieval independently. Five mutually exclusive operational outcomes are derived within the 20,000-row AmazonQAC test file and evaluated with a strict session-disjoint temporal split; product-query provenance is not used as a route class. Across five seeds, the combined word/character TF-IDF router reached 0.473 mean accuracy (SD 0.031) and 0.330 mean macro-F1 (SD 0.018), indicating uneven performance across routes. On the 951-row temporal test set, context-nearest completion achieved 0.033 Success@10 and 0.025 MRR@10, compared with 0.029 and 0.023 for confidence-gated routing. The paired MRR@10 difference was 0.002 (95% CI -0.007 to 0.010; Holm-adjusted p = 1.000). Product retrieval on 480 WANDS queries and 42,994 products produced a different pattern: a word/character lexical hybrid reached 0.690 nDCG@10 and 0.504 Exact-MRR@10, significantly exceeding the title-only baseline on both measures. Route-aware decision-making is therefore useful as an architectural separation of entry actions, but it does not automatically improve exact autocomplete ranking over a strong session-context baseline.

Keywords: *Query Autocomplete, Session Context, Search-Entry Routing, Lightweight Lexical Classification*

I. INTRODUCTION

Modern search entry points are no longer passive text boxes. In web and e-commerce systems, the entry layer decides whether to continue suggesting completions, submit the current text, use recent session context, or fall back to a broader search. Product relevance is a related downstream task, but it is not itself an operational completion route. This distinction is consistent with classic information-retrieval views of navigational, informational, and transactional intent (Broder, 2002; Rose & Levinson, 2004). In e-commerce, the same interface connects to product classification, personalized recommendation, and next-purchase prediction, where compact user and item representations can turn short behavioral traces into product decisions (Lu & Zhou, 2023; K. Xu et al., 2024; Xin, 2026b). Query-log research likewise shows that sequential behavior contains information that a single short query cannot expose (Joachims, 2002; Radlinski & Joachims, 2005).

Query autocomplete predicts the final query before submission, but many prefixes are short, underspecified, or ambiguous. Context-sensitive query auto-completion can therefore use recent queries to improve candidate selection (Bar-Yossef & Kraus, 2011), while broader surveys identify prefix matching, candidate generation, ranking, personalization, and latency as the core design dimensions (Cai & de Rijke, 2016). Large models expand the available semantic signals,

Received on December, 2026; Revised on January, 2026; Accepted on February, 2026; Published on April, 2026

and teacher-student knowledge distillation can transfer such behavior when teacher outputs and a defined transfer objective are available (Hinton et al., 2015; X. Xu et al., 2024). The present study does not implement that protocol: it evaluates a transparent TF-IDF and linear-classification baseline over observed query-log fields. Evidence controls remain relevant because generated or reranked suggestions can detach from catalog, log, or session evidence; lightweight hallucination firewalls and evidence-chain RAG studies offer useful patterns for source consistency and provenance (C. Li et al., 2023, 2024; Jin, 2025a).

This study asks two questions. First, how reliably can a low-latency lexical classifier distinguish five operational autocomplete outcomes under source-controlled, session-disjoint temporal validation, and does confidence-gated routing improve exact completion ranking over context-nearest retrieval? Second, how well do simple lexical product retrievers perform on independently judged customer queries? AmazonQAC provides naturalistic prefixes, final search terms, timestamps, session identifiers, and past search terms (Everaert et al., 2024). WANDS supplies historical customer queries, a full product catalog, and human relevance judgments (Chen et al., 2022). The metadata-generated Amazon Search Benchmark serves only as a contextual contrast between synthetic query generation and independent relevance assessment; it is not used to define routes or produce the reported metrics (ApexLearningCurve, 2024).

The paper makes four contributions. It formalizes five mutually exclusive post-hoc operational routes from a single query environment, including normalization, thresholds, conflict priority, and examples. It evaluates word, character, and combined TF-IDF representations with a class-balanced linear classifier across repeated seeds and reports class-level error patterns. It compares exact autocomplete strategies with session-cluster confidence intervals and paired tests. Finally, it evaluates title, metadata, and hybrid lexical retrieval on WANDS with query-bootstrap intervals and paired randomization tests. Strong lexical baselines remain important for fast and interpretable retrieval (Robertson & Zaragoza, 2009), and the route-first decomposition parallels broader work that separates action selection, failure analysis, and confidence before downstream execution (C. Li et al., 2026; Zhong et al., 2026).

The results separate architectural usefulness from ranking effectiveness. The combined router improves over word-only and character-only models, but its five-seed macro-F1 is 0.330 and the context-aware route remains particularly difficult. Context-nearest completion obtains the highest exact MRR@10, while confidence-gated routing is slightly lower and not significantly different. Candidate coverage is below 6.4% even for the shortest prefix bucket after context augmentation, which limits all retrieval-only completion methods. On WANDS, richer lexical representations significantly improve ordering quality, although Recall@10 and Hit@10 gains are not statistically

reliable. These findings support modular routing and retrieval, while cautioning against treating a route classifier as an automatic improvement to exact autocomplete ranking.

II. LITERATURE REVIEW

Search intent has long been treated as a central variable in retrieval systems. Broder (2002) organized web search around broad intent classes, and Rose and Levinson (2004) refined the idea that search goals shape system behavior. In an e-commerce entry box, these classic categories translate into operational routing decisions: a complete product query can move directly to retrieval, an incomplete prefix needs autocomplete, and a repeated session topic benefits from context. Clickthrough and query-chain work also established that search behavior is sequential, not isolated (Joachims, 2002; Radlinski & Joachims, 2005).

Query autocomplete research builds on this sequential view. Bar-Yossef and Kraus (2011) showed that recent queries improve completions for short prefixes, and Cai and de Rijke (2016) summarized the field around candidate generation, ranking, and context. AmazonQAC strengthened this line of research by providing a naturalistic resource with prefixes, final search terms, timestamps, session identifiers, and past search terms (Everaert et al., 2024). Its official 20,000-row test period follows the much larger training period, enabling temporal evaluation rather than a random split of repeated sessions (Amazon, 2024). Realistic evaluation therefore requires both the prefix-to-final-query relation and careful control of session dependence.

The second literature stream is product-aware search. Traditional lexical retrieval methods, especially BM25, remain strong because short queries often share exact tokens with titles, categories, or product attributes (Robertson & Zaragoza, 2009). Language-model approaches to retrieval connect query likelihood and document text in a way that remains useful for product matching (Ponte & Croft, 1998), while neural retrieval and transformer-based ranking learn richer matching functions (Devlin et al., 2019; Reimers & Gurevych, 2019). Retrieval-augmented generation further showed the value of grounding generation in retrieved evidence (Lewis et al., 2020), and product-aware query completion applies catalog grounding to e-commerce suggestions (Sun et al., 2024). WANDS complements these methods with historical customer queries and human relevance judgments for Exact, Partial, and Irrelevant matches (Chen et al., 2022). Recent e-commerce studies extend retrieval into customer representation, product classification, recommendation, reranking, and explanation (Lu & Zhou, 2023; K. Xu et al., 2024; Xin, 2026b; Chang et al., 2023; X. Chang et al., 2026; Xin, 2026a). Together, this work motivates evaluating product retrieval separately from operational autocomplete routing.

Product search also has a user-interface dimension. LLM-style recommendation cards and scientific-search evidence cards both emphasize that ranked results should expose evidence

hierarchy, source attribution, and confidence rather than only outputting a final recommendation (B. Zhang et al., 2025; Jin, 2025b). For a search entry box, the analogous interface need not display a full explanation at every keystroke, but its logs and debugging tools should retain the route evidence, candidate source, and confidence used to produce a suggestion.

A third literature stream concerns model compression and distillation. Hinton et al. (2015) formalized the teacher-student framework for transferring knowledge from a larger model to a smaller one. Transformer research created powerful encoders and sequence models (Vaswani et al., 2017; Devlin et al., 2019), and their cost motivated compact alternatives. DistilBERT used distillation during pretraining (Sanh et al., 2019), TinyBERT transferred transformer behavior at general and task-specific stages (Jiao et al., 2020), and compact-model work showed that pretraining remains valuable for small students (Turc et al., 2019). Surveys now cover the transfer of reasoning, alignment, and domain behavior from large language models (X. Xu et al., 2024), while production research treats cost, latency, and workload placement as first-class constraints (Xin, 2025). Accordingly, the TF-IDF linear model is described as a lexical classifier because it receives no teacher outputs.

Query understanding is a plausible target for true distillation because queries are short and latency-sensitive. QUILL combined retrieval augmentation with multi-stage distillation for query-intent understanding (Srinivasan et al., 2022). Here, the research question is narrower: how far can an efficient lexical router go when it is trained only on operational labels derived within AmazonQAC? The Amazon Search Benchmark offers paired short and long queries generated from target product metadata (ApexLearningCurve, 2024), but that provenance makes it better suited to illustrating synthetic-query conditions than to supervising an operational route. Work on conversational text-to-SQL and code intelligence similarly separates evidence retrieval, task feedback, and explanation, supporting route choice as an observable intermediate decision (Y. Li, 2023, 2024).

The unresolved connection among these streams is at the search-entry boundary. Query autocomplete predicts the final query, product retrieval ranks catalog items for a submitted query, and model compression reduces inference cost. IntentRouter-QAC treats route selection, completion ranking, and product retrieval as separate measurable layers. The router distinguishes five operational outcomes within one query environment; autocomplete methods use prefix and session evidence; and WANDS provides an independent product-retrieval test. This separation prevents a model from treating dataset provenance as a semantic class.

Reliability-oriented RAG work also motivates fallback and confidence gating. Evidence-calibrated unanswerable QA, budgeted multi-hop retrieval, trust-calibrated multilingual RAG,

and ambiguity-aware log-anomaly narratives all prefer abstention or selective action when retrieval coverage is weak or evidence is inconsistent (Y. Chen & Xu, 2026; Nie & Zheng, 2023; Su et al., 2025; Zhong, Chen, et al., 2025). FALLBACK_SEARCH is not a safety refusal in this study; it is an operational outcome indicating that the submitted query no longer follows the observed prefix. Confidence gating applies the broader principle conservatively by allowing route control only above a development-selected threshold.

Latency and interpretability motivate a shallow lexical classifier at the entry point. The search box receives every keystroke, so its first decision must be made with very little text and a bounded response time. Linear classifiers over word and character n-grams provide efficient and inspectable text decisions (Joulin et al., 2017). Unlike DistilBERT and TinyBERT, which are neural students trained with distillation objectives (Jiao et al., 2020; Sanh et al., 2019), the model evaluated here is neither generative nor teacher-trained. A transformer reranker could still be added after candidate generation without changing that first-stage definition.

Session context can improve short-prefix ranking while raising privacy and evaluation concerns. Federated topic-preference learning and privacy-robust incrementality estimation suggest ways to limit raw-history exposure and to evaluate decisions made from biased behavioral logs (Bai et al., 2026; Kuo et al., 2025). In this study, recent search terms are used only as bounded within-session features; no persistent user profile is constructed. The remaining safeguards and limitations of timestamped query sequences are specified in the methods.

III. RESEARCH METHOD

Table 1 summarizes the three public resources considered and their distinct roles. Measured autocomplete and route results come from the official 20,000-row AmazonQAC test file, whose seven fields are query_id, session_id, past_search_terms, prefix, prefix_typed_time, final_search_term, and search_time (Amazon, 2024; Everaert et al., 2024). Product retrieval is evaluated on WANDS, which contains 480 historical customer queries, 42,994 products, and 233,448 relevance judgments (Chen et al., 2022). The Amazon Search Benchmark is used only to characterize the easier target-derived query condition discussed in the limitations (ApexLearningCurve, 2024). Text normalization applies Unicode NFKC, case folding, replacement of non-alphanumeric characters with spaces, and whitespace collapse.

AmazonQAC is distributed as de-identified behavioral data; its data card reports regex- and model-based removal of personally identifiable information and a minimum recurrence rule for retained search terms (Amazon, 2024). Session identifiers are used only to group and order events, timestamps only to enforce chronology, and at most five recent terms are retained for each row. No attempt is made to identify users or construct cross-session profiles. One isolated normalized

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418-440
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

query transition is reported for each labeling rule without session identifiers; all other results are aggregated. Timestamped histories can nevertheless remain sensitive because repeated sequences and quasi-identifiers may permit linkage across records (Pàmies-Estrems & Garcia-Alfaro, 2023).

Table 1. Dataset inventory and experimental role

Dataset	Scale	Field Used	Role In This Study
AmazonQAC	20.000 query-completion events	Prefix, final query, up to five recent terms, session/time fields	Five-route construction, candidate coverage, and exact QAC
WANDS	480 queries, 42.994 products, 233.448 judgments	Query, title, product class, category hierarchy, relevance	Independent product retrieval with human judgments
Amazon Search Benchmark	20.373 product records	Short query, long query, target product text	Scope comparison for metadata-generated queries, excluded from route labels

A strict session-disjoint temporal partition was constructed within the October 1-14 file. Sessions ending before October 11 form training; sessions wholly contained on October 11-12 form development; and sessions beginning on or after October 13 form testing. Sessions crossing either boundary were excluded so that no identifier appears in more than one partition. As Table 2 reports, this yields 7,223 training rows from 3,425 sessions, 1,329 development rows from 852 sessions, and 951 test rows from 640 sessions; 10,497 boundary-crossing rows are excluded. The intersection of session identifiers among retained partitions is zero.

Table 2. AmazonQAC temporal split and full-file statistics

Scope	Rows	Sessions	UTC period / description
Full file	20,000	6,679	Oct. 1-14, 2023; 16,667 final queries; 14,921 prefixes
Training	7,223	3,425	Oct. 1 00:11-Oct. 10 23:58; sessions end before Oct. 11
Development	1,329	852	Oct. 11 00:04-Oct. 12 23:57; sessions contained in window
Test	951	640	Oct. 13 00:03-Oct. 14 23:59; sessions begin on/after Oct. 13
Boundary-crossing	10,497	1,762	Excluded to preserve chronology and zero session overlap

Product retrieval uses all 480 WANDS queries against the full 42,994-product catalog. Repeated query-product judgments are collapsed by retaining the maximum relevance gain; this converts 233,448 raw labels into 231,873 unique judged pairs and resolves 14 conflicting repeats.

Table 3 gives the WANDS statistics used in the evaluation.

Statistic	Value
Queries	480
Products	42,994
Raw relevance judgments	233,448
Unique judged query-product pairs	231,873
Raw Exact / Partial / Irrelevant labels	25,614 / 146,633 / 61,201
Query classes / product classes	188 / 860
Category hierarchies	1,623
Mean normalized query tokens	3.381
Conflicting repeated pairs	14; resolved by maximum gain

WANDS relevance gains are 2 for Exact, 1 for Partial, and 0 for Irrelevant. Exact or Partial labels count as relevant for Recall@10 and Hit@10, whereas Exact alone defines Exact-MRR@10. Unjudged query-product pairs receive gain zero. These choices follow the three-level WANDS

judgments and preserve the distinction between retrieving a broadly relevant item and ranking an exact match first (Chen et al., 2022).

Table 4. Operational route rules, examples, and training distribution

Route (priority)	Deterministic condition	Training n (%)	Normalized training transition
DIRECT_SEARCH (1)	f starts with p; p=f, or r≥0.80 and ≤3 characters remain	703 (9.7%)	dishwasher → dishwasher
CONTEXT_AWARE_SUGGEST (2)	Non-exact f; f/h containment (shorter≥4 characters) or token overlap≥0.67	1,109 (15.4%)	large travel → large travel backpack for men; recent: travel backpack
FALLBACK_SEARCH (3)	f does not start with p	1,336 (18.5%)	halloween wil → halloween hanging ghost
AUTOSUGGEST (4)	Remaining prefix-consistent row; |p|≤4 or r≤0.60	3,513 (48.6%)	wo → womens scrunch boot socks
REFINEMENT_SUGGEST (5)	All remaining prefix-consistent rows	562 (7.8%)	womens cowboy → womens cowboy boots

Five mutually exclusive post-hoc operational outcome labels are derived from the normalized prefix (p), final query (f), and up to five recent browsing history terms (H). The prefix ratio is defined as $r = |p| / |f|$, while token overlap between the final query and a history term $h$is calculated as the size of the token intersection divided by the size of the smaller token set. The labeling procedure follows the priority order summarized in Table 4. An interaction is labeled DIRECT_SEARCH when the final query begins with the prefix and either exactly matches it ($p = f$) or satisfies both $r \geq 0.80$and a maximum extension of three additional characters. If this condition is not met, the interaction is assigned CONTEXT_AWARE_SUGGEST when the final query differs from the prefix and exhibits a strong contextual relationship with at least one history term.

Specifically, this relationship is established when one string contains the other and the shorter string is at least four characters long, or when the token overlap is at least 0.67, provided that both the final query and the history term contain at least two tokens. Interactions in which the final query does not begin with the prefix are classified as FALLBACK_SEARCH. For the remaining prefix-consistent interactions, AUTOSUGGEST is assigned when the prefix length does not exceed four characters or the prefix ratio is no greater than 0.60. All remaining cases are labeled REFINEMENT_SUGGEST. When browsing history is unavailable, the token-overlap value is defined as zero. Because the decision rules are evaluated sequentially, the first satisfied condition uniquely determines the final outcome label, ensuring deterministic conflict resolution.

The corresponding pseudocode is as follows. First, the prefix (p), final query (f), and browsing history (H) are normalized, and the prefix ratio is computed as $r = |p| / |f|$. If the final query begins with the prefix and either $p = f$or $(r \geq 0.80 \wedge |f| - |p| \leq 3)$, the interaction is labeled

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418-440
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

DIRECT_SEARCH. Otherwise, if the final query differs from the prefix and satisfies the strong_context_match criterion with the browsing history, the interaction is assigned CONTEXT_AWARE_SUGGEST. If the final query does not begin with the prefix, the interaction is classified as FALLBACK_SEARCH. When the prefix length is at most four characters or the prefix ratio is no greater than 0.60, the interaction is labeled AUTOSUGGEST. All remaining interactions are assigned REFINEMENT_SUGGEST.

The function strong_context_match implements the containment and token-overlap criteria defined previously, while the ordered sequence of decision rules ensures deterministic conflict resolution when multiple conditions could apply. Importantly, the final query is used exclusively for retrospective label generation and is never included as an input feature during model training or inference, thereby preventing target leakage. Table 4 summarizes the complete training distribution together with one representative normalized training transition for each routing rule. Finally, product-query provenance is intentionally excluded as a routing label because combining source identity with operational outcomes could encourage the classifier to exploit source-specific lexical correlations rather than learning the intended routing behavior (Wang & Culotta, 2020). Instead, product relevance is evaluated independently through the WANDS experiment.

Four routing classifiers are evaluated to assess the effectiveness of the proposed query-routing framework. The first is a deterministic heuristic baseline that relies on four handcrafted features: the number of training candidates, the maximum prefix–history overlap, prefix length, and prefix token count. The second model is a word-level representation based on TF–IDF word unigrams and bigrams, with a vocabulary limited to 8,000 fitted features. The third employs a character-level TF–IDF representation using within-word character 3–5-grams, with a maximum of 12,000 fitted features. The proposed IntentRouter-QAC integrates both lexical representations by concatenating them with six additional numerical features: the logarithm of prefix length, prefix token count, the logarithm of context count, maximum prefix–history token overlap, a binary prefix–history substring indicator, and the logarithm of the training-candidate count.

To prevent data leakage, all feature representations are fitted exclusively on the training partition before being applied to the validation and test sets. Each feature representation is subsequently classified using a class-balanced linear stochastic gradient descent (SGD) logistic regression model with L2 regularization, a regularization coefficient of $\alpha = 10^{-5}$, a maximum of 5,000 optimization iterations, and a convergence tolerance of 10^{-4}. Model stability is assessed using five independent random seeds (13, 23, 42, 73, and 101) to quantify fitting variability. As summarized in Table 4, evaluation emphasizes both macro-F1 and per-route performance, since overall accuracy alone may obscure deficiencies in minority routing classes, particularly under imbalanced class distributions (Y. Chen et al., 2023; Jin et al., 2024a, 2024b; Zhong et al., 2023).

Autocomplete is evaluated on the 951-row AmazonQAC test partition with Success@1, Success@5, Success@10, MRR@10, and top-1 token F1. Echo prefix returns the typed prefix; global popularity returns frequent training outcomes; and prefix popularity retrieves training outcomes that begin with the prefix. Context-nearest completion adds recent terms that begin with the prefix to the prefix-consistent training pool and ranks each candidate q by log(1 + training frequency(q)) + 1.50 times its maximum token-overlap coefficient with a recent term + 0.40 times a containment indicator; ties are resolved by training frequency and then alphabetically. At route confidence below 0.50, the confidence-gated method defaults to context-nearest completion. At or above 0.50, DIRECT_SEARCH prepends the typed prefix to prefix-popularity candidates, CONTEXT_AWARE_SUGGEST uses context-nearest completion, FALLBACK_SEARCH places the most recent terms before globally popular outcomes, and AUTOSUGGEST or REFINEMENT_SUGGEST uses prefix popularity. Threshold 0.50 maximized development MRR@10. Table 5 reports whether the gold final query is available before ranking, both in the training-query pool and after recent-context augmentation.

Uncertainty is estimated at the dependency level. AmazonQAC confidence intervals use 2,000 session-cluster bootstrap samples, and paired autocomplete comparisons use 20,000 two-sided session-cluster permutations with Holm correction. Route models are also summarized across five fixed seeds. WANDS intervals use 10,000 query bootstrap samples, and model differences use 100,000 paired query-level sign-flip permutations with Holm correction within each metric. This choice of paired and clustered units follows guidance that statistical tests should match both the evaluation measure and the dependence structure (Dror et al., 2018). The log-based autocomplete comparison remains an offline benchmark rather than a causal estimate of user utility; policy changes would require conservative off-policy or incrementality evaluation (Bai et al., 2026; Lu et al., 2024; Mu et al., 2023, 2025; Ye et al., 2026).

Table 5. Candidate coverage by normalized prefix length

Prefix characters	Rows	Training-query coverage	Context-augmented coverage	Mean training candidates	Mean augmented candidates
1-4	299	0.033	0.064	99.328	99.559
5-8	225	0.018	0.036	3.524	3.667
9-12	172	0.012	0.017	1.087	1.163
13-16	125	0.008	0.016	0.048	0.224
17+	130	0.000	0.015	0.015	0.169

For WANDS retrieval, title-only TF-IDF uses word unigrams and bigrams; the metadata index adds product class and category hierarchy; and the hybrid score combines the metadata-word score (weight 0.70) with title character 3-5-gram similarity (weight 0.30). Metrics are nDCG@10, Recall@10,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418-440
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Exact-MRR@10, Hit@10, and single-process latency after index construction. Figure 1 shows the complete pipeline, with AmazonQAC operational routing separated from WANDS product relevance.

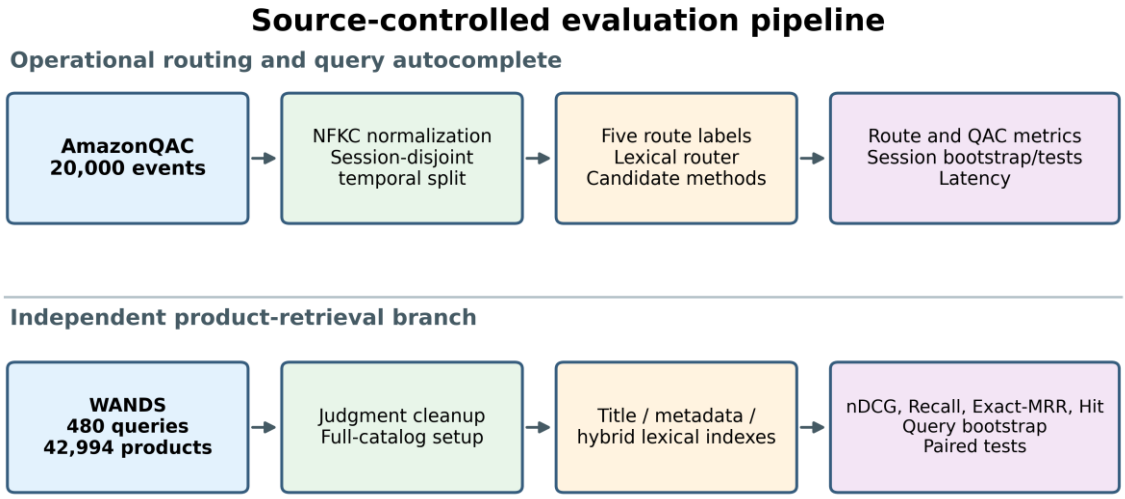

Figure 1. Experimental pipeline separating AmazonQAC route and autocomplete evaluation from WANDS product retrieval

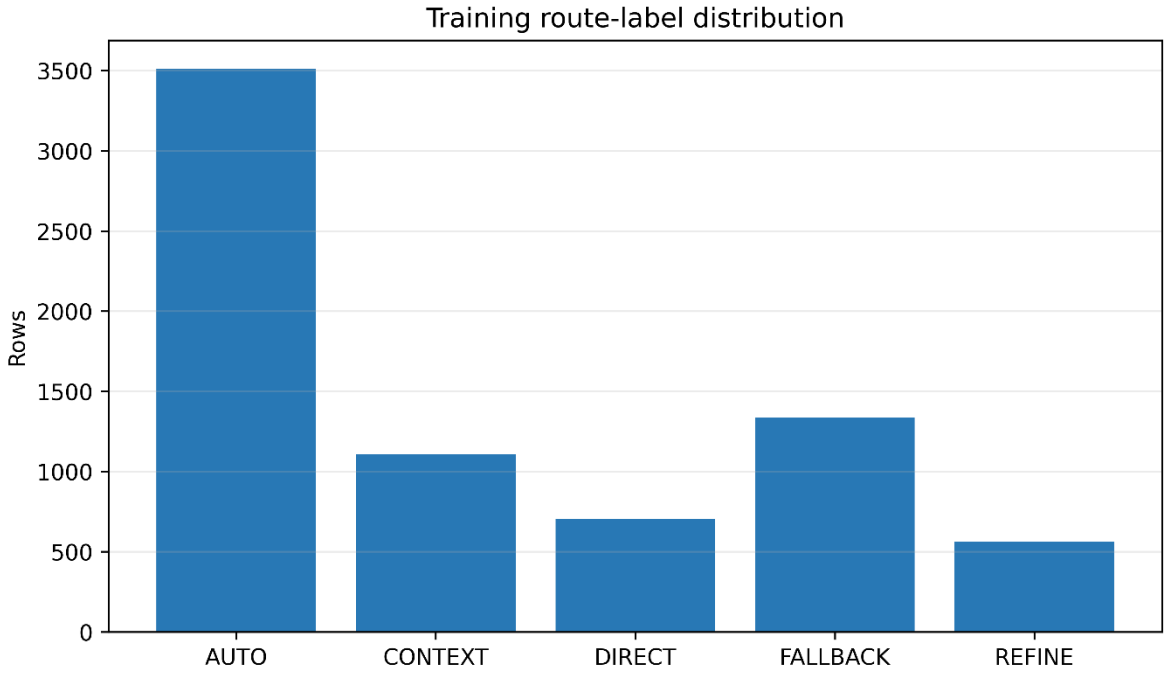

Figure 2. Training distribution of the five AmazonQAC operational routes

IV. RESULT AND DISCUSSION

A. Result

The full AmazonQAC file contains 16,667 unique normalized final search terms and 14,921 unique normalized prefixes. The mean prefix length is 9.20 characters, the mean final-query length is 3.34 tokens, and the mean retained context is 3.92 recent terms; 76.2% of final queries

start with the prefix and 3.84% are already complete. Table 2 adds the exact temporal split and session counts. Before duplicate resolution, the 233,448 WANDS judgments comprised 25,614 Exact, 146,633 Partial, and 61,201 Irrelevant labels; maximum-gain resolution yielded 231,873 unique query-product pairs, as summarized in Table 3. Table 4 shows a materially imbalanced route distribution: AUTOSUGGEST accounts for 3,513 of 7,223 training rows (48.6%), whereas REFINEMENT_SUGGEST accounts for 562 (7.8%). Figure 2 visualizes all five route counts. Every route comes from the same AmazonQAC environment, and product-query provenance is not defined as a class.

Table 6. Route classification performance on the AmazonQAC test partition

Model	Accuracy	Macro-F1	Weighted-F1
Heuristic (deterministic)	0.416 [0.383, 0.450]	0.239 [0.217, 0.262]	0.408 [0.375, 0.442]
Word n-gram, seed 42	0.334 [0.302, 0.365]	0.207 [0.180, 0.234]	0.341 [0.309, 0.374]
Character n-gram, seed 42	0.352 [0.319, 0.387]	0.229 [0.201, 0.259]	0.360 [0.326, 0.397]
Combined lexical, seed 42	0.455 [0.423, 0.489]	0.320 [0.287, 0.354]	0.472 [0.438, 0.507]
Combined lexical, five seeds	0.473 ± 0.031	0.330 ± 0.018	0.482 ± 0.017

Table 6 reports both primary-seed confidence intervals and five-seed variability. At seed 42, the combined lexical router achieved 0.455 accuracy (95% CI 0.423-0.489), 0.320 macro-F1 (0.287-0.354), and 0.472 weighted-F1 (0.438-0.507). Across five seeds, its means were 0.473 accuracy (SD 0.031), 0.330 macro-F1 (SD 0.018), and 0.482 weighted-F1 (SD 0.017). These values exceed the word-only, character-only, and deterministic baselines, but the macro-F1 remains too low to describe five-route prediction as broadly reliable.

Table 7. Per-route precision, recall, and F1 for the combined lexical router

Route	Precision	Recall	F1	Support
Autosuggest	0.763	0.625	0.687	499
Context-aware suggest	0.444	0.121	0.190	132
Direct search	0.194	0.313	0.240	67
Fallback search	0.225	0.371	0.280	175
Refinement suggest	0.174	0.244	0.203	78

Table 7 shows substantial variation by route. AUTOSUGGEST has the highest F1 at 0.687, while CONTEXT_AWARE_SUGGEST has the lowest at 0.190; DIRECT_SEARCH, FALLBACK_SEARCH, and REFINEMENT_SUGGEST obtain F1 values of 0.240, 0.280, and 0.203. Figure 3 shows the confusion matrix. The largest off-diagonal errors are AUTOSUGGEST to FALLBACK_SEARCH (118 rows), CONTEXT_AWARE_SUGGEST to FALLBACK_SEARCH (50), and FALLBACK_SEARCH to AUTOSUGGEST (42). The pattern suggests that prefix and shallow history features often cannot separate behaviorally defined actions that share similar surface forms.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418-440
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

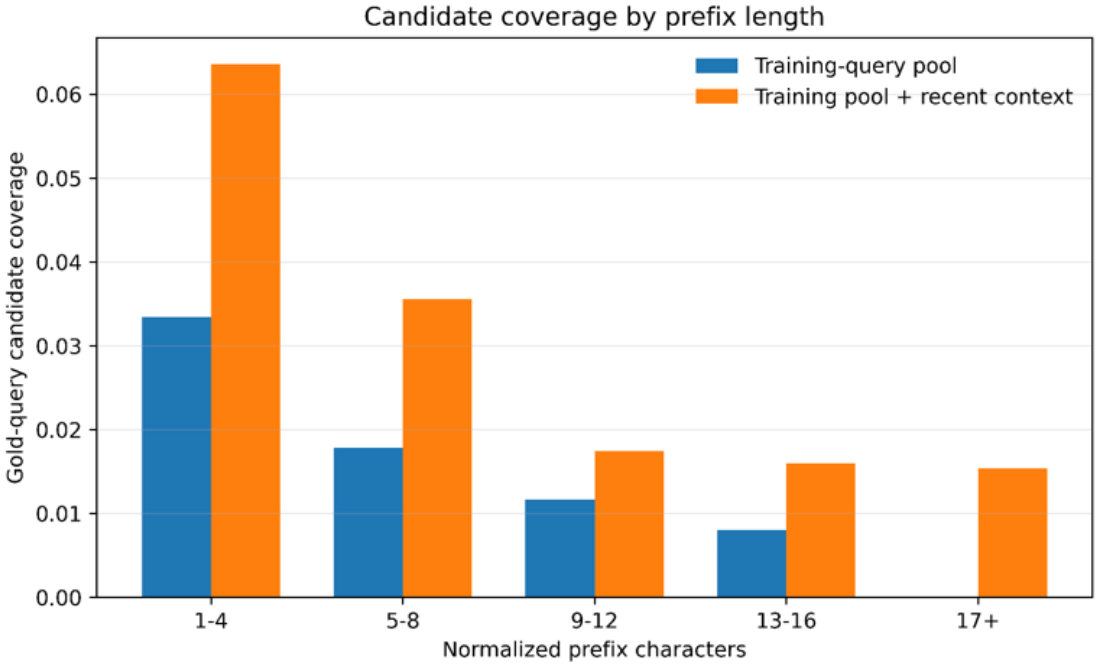

Figure 3. Confusion matrix for the combined lexical router on the 951-row test partition

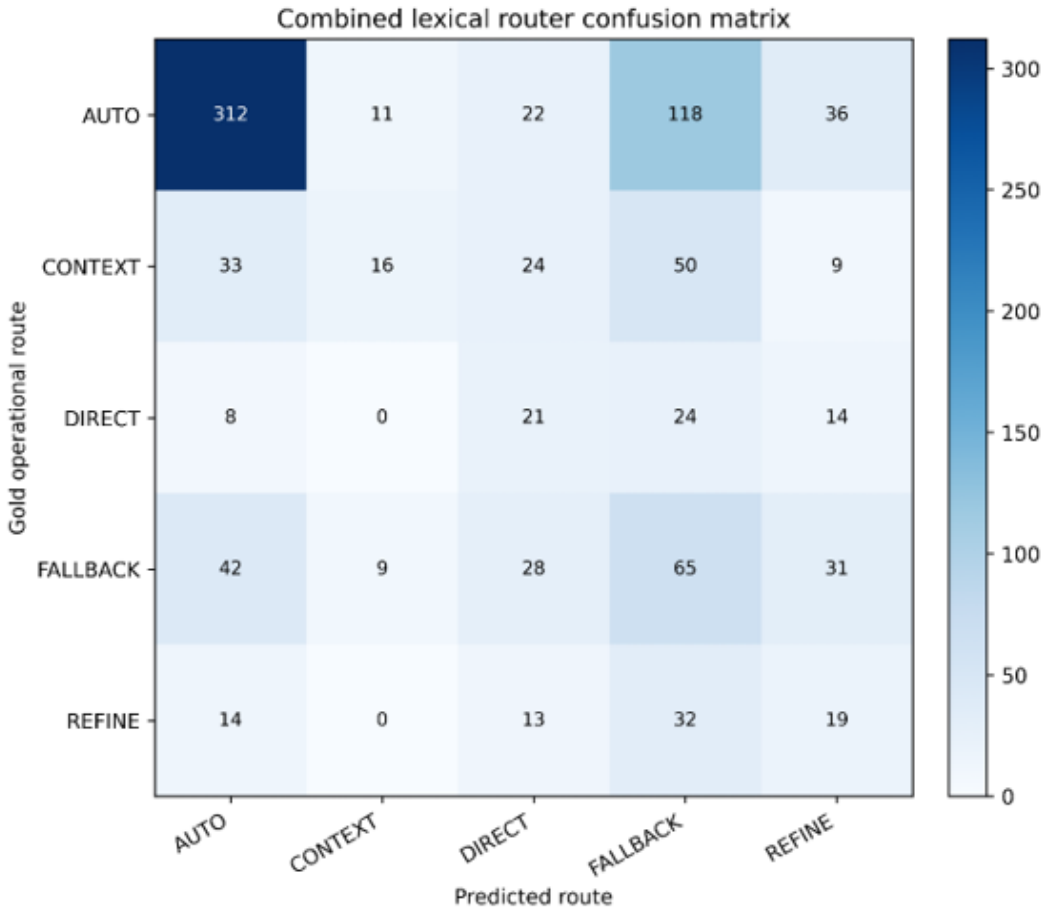

Figure 4. Candidate coverage before and after recent-context augmentation

Candidate availability is a severe constraint on exact autocomplete. As Table 5 reports, the training-query pool contains the gold final query for 3.34% of 1-4 character prefixes, and recent context raises coverage to 6.35%. For the 17+ bucket, base coverage is zero and context-augmented coverage is 1.54%. Figure 4 shows that both sources decline with prefix length, while the context pool provides a consistent but small absolute gain. These values are computed only from the temporally prior training partition, so unseen and shifted queries remain unrecoverable by a retrieval-only candidate generator.

Table 8. Autocomplete results on the AmazonQAC temporal test partition

Method	Success@1	Success@5	Success@10	MRR@10	Top-1 token F1	Latency (ms/query)
Echo prefix	0.021	0.021	0.021	0.021	0.336	0.0002
Global popularity	0.000	0.000	0.001	0.0001	0.006	0.0002
Prefix popularity	0.006	0.009	0.013	0.008	0.072	0.0098
Context-nearest	0.022	0.029	0.033	0.025	0.134	0.1998
Confidence-gated route	0.021	0.026	0.029	0.023	0.184	0.1410

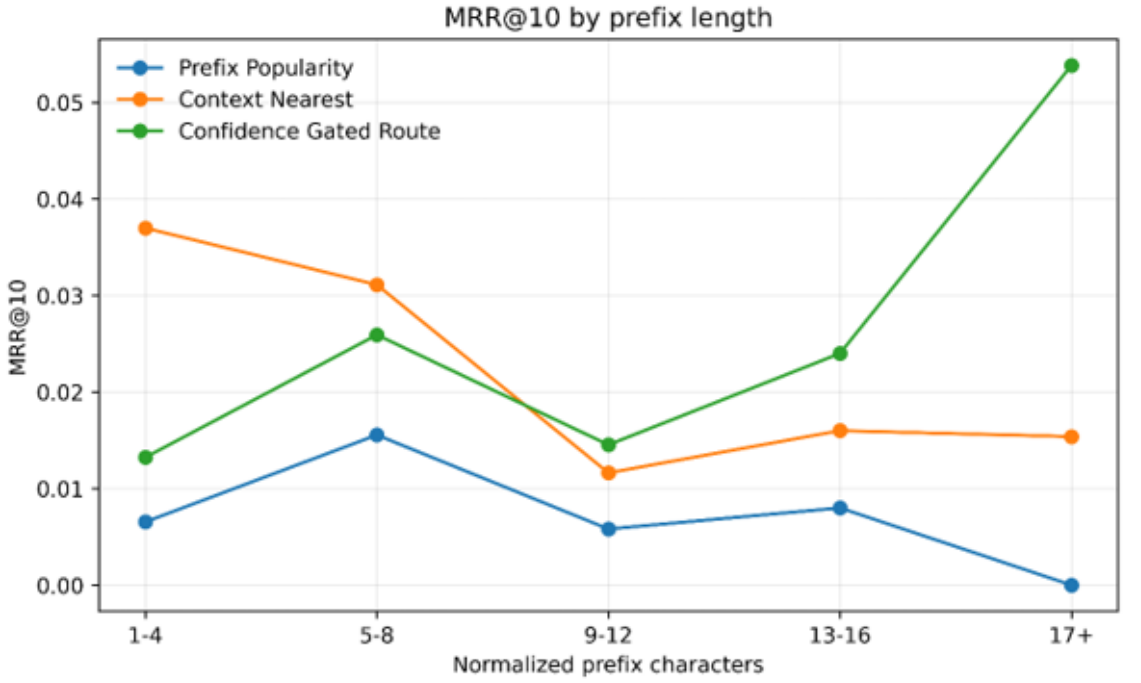

Figure 5. Autocomplete Success@10 and MRR@10 on the AmazonQAC test partition

Table 8 and Figure 5 compare the five completion methods. Context-nearest completion achieves the highest Success@10, 0.0326 (95% CI 0.0215-0.0445), and the highest MRR@10, 0.0253 (0.0155-0.0367). Confidence-gated routing reaches 0.0294 Success@10 (0.0191-0.0406) and 0.0234 MRR@10 (0.0147-0.0331). Prefix popularity is lower at 0.0126 Success@10 and 0.0078 MRR@10. The gated method has higher top-1 token F1 than context-nearest (0.184 versus 0.134), but it does not improve exact final-query ranking.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418-440
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

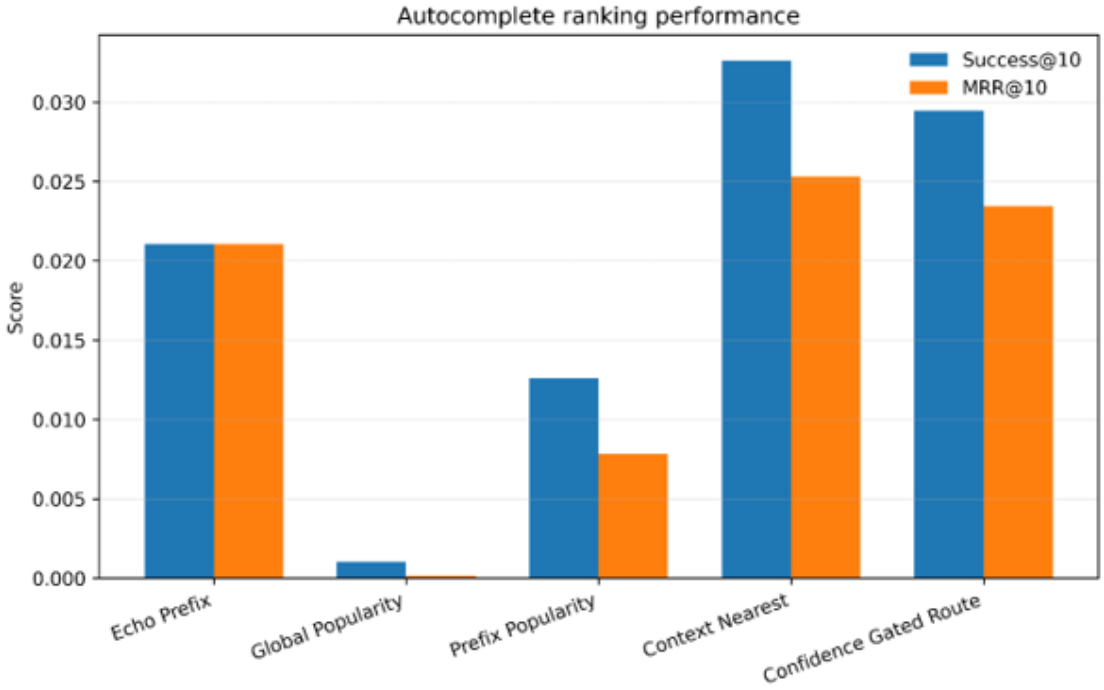

Figure 6. Autocomplete MRR@10 by normalized prefix length

Table 9. Autocomplete MRR@10 by normalized prefix length

Prefix characters	Rows	Echo prefix	Prefix popularity	Context-nearest	Confidence-gated route
1-4	299	0.003	0.007	0.037	0.013
5-8	225	0.009	0.016	0.031	0.026
9-12	172	0.006	0.006	0.012	0.015
13-16	125	0.056	0.008	0.016	0.024
17+	130	0.069	0.000	0.015	0.054

Table 9 breaks MRR@10 down by prefix length, and Figure 6 displays the same comparison. Context-nearest is strongest in the 1-4 and 5-8 buckets, with MRR@10 of 0.0370 and 0.0311. Confidence-gated routing is lower for short prefixes but reaches 0.0538 in the 17+ bucket, where only seven exact matches occur among 130 rows. The bucket result should therefore not be generalized beyond this test partition.

Table 10. Ablation of session context and confidence-gated route control

Variant	Context used	Route control	Success@10	MRR@10	Top-1 token F1	Threshold
Prefix popularity	No	No	0.013	0.008	0.072	-
Context-nearest	Yes	No	0.033	0.025	0.134	-
Confidence-gated route	Yes	Yes	0.029	0.023	0.184	0.50

Table 10 isolates the effects of context and route control. Adding recent context to prefix popularity increases MRR@10 from 0.0078 to 0.0253 and Success@10 from 0.0126 to 0.0326. Adding

confidence-gated route decisions reduces these values to 0.0234 and 0.0294. The paired context-nearest minus gated difference is 0.0019 MRR@10 (95% CI -0.0065 to 0.0102; Holm-adjusted p = 1.000) and 0.0032 Success@10 (-0.0055 to 0.0119; Holm-adjusted p = 1.000). The experiment therefore supports context as the primary exact-ranking signal, while treating the router as a conditional action layer. This conservative policy is consistent with uncertainty-aware and off-policy selection principles (Lu et al., 2024; Ye et al., 2026).

Table 11. Product retrieval results on WANDS

Model	nDCG@10 [95% CI]	Recall@10 [95% CI]	Exact-MRR@10 [95% CI]	Hit@10 [95% CI]	Mean / P95 latency (ms)
Title-only word TF-IDF	0.640 [0.612, 0.669]	0.056 [0.050, 0.063]	0.441 [0.401, 0.483]	0.931 [0.906, 0.954]	2.905 / 3.410
Title + class/hierarchy	0.656 [0.626, 0.685]	0.056 [0.050, 0.062]	0.479 [0.438, 0.522]	0.938 [0.915, 0.958]	4.284 / 5.290
Word/character hybrid	0.690 [0.663, 0.717]	0.059 [0.053, 0.065]	0.504 [0.463, 0.545]	0.950 [0.929, 0.969]	15.866 / 18.710

Table 11 reports product retrieval over all 480 WANDS queries and the full catalog. The word/character hybrid obtains 0.690 nDCG@10 (95% CI 0.663-0.717), 0.059 Recall@10 (0.053-0.065), 0.504 Exact-MRR@10 (0.463-0.545), and 0.950 Hit@10 (0.929-0.969). Figure 7 compares its ordering metrics with the two word-only indexes. Relative to title-only retrieval, the hybrid improves nDCG@10 by 0.0498 and Exact-MRR@10 by 0.0624; both differences remain significant after Holm correction (p = 0.00003). Relative to metadata-word retrieval, its gains are 0.0342 nDCG@10 (Holm-adjusted p = 0.00003) and 0.0249 Exact-MRR@10 (Holm-adjusted p = 0.00103). Recall@10 and Hit@10 differences are not significant, so the evidence supports improved ordering rather than a reliable coverage gain.

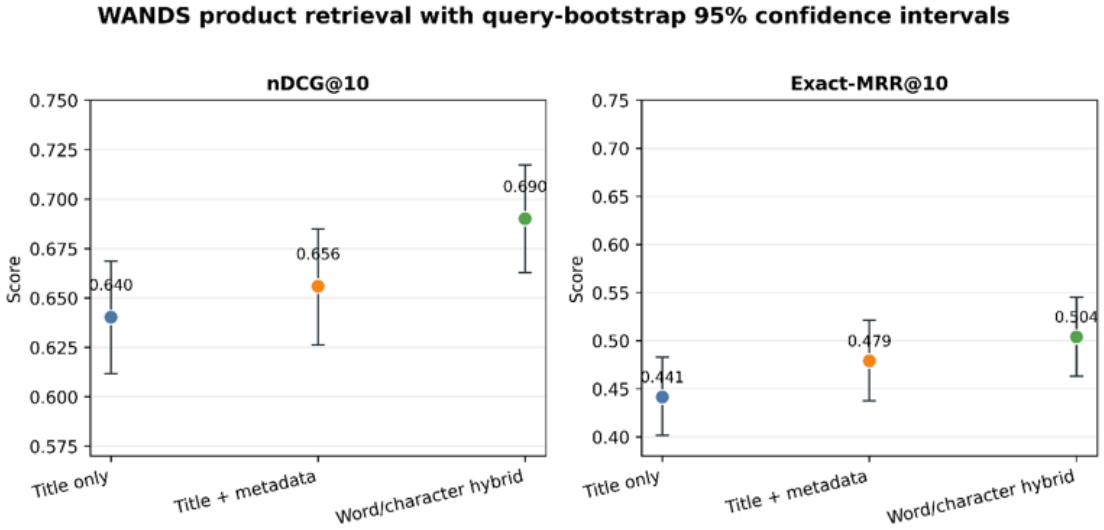

Figure 7. WANDS lexical-retrieval results with query-bootstrap 95% confidence intervals

Table 12. Runtime and representation-size summary

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 418-440
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Component	Rows / queries	Primary metric	Value	Mean latency (P95 in parentheses for WANDS; ms/query)	Model or index size
Combined route classifier	951	Accuracy	0.455	0.1276	20,006 input features
Prefix-popularity QAC	951	MRR@10	0.008	0.0098	6,161 indexed outcomes
Context-nearest QAC	951	MRR@10	0.025	0.1998	6,161 outcomes + up to five recent terms
Confidence-gated QAC	951	MRR@10	0.023	0.1410	20,006 features + 6,161 outcomes
WANDS title-only	480	nDCG@10	0.640	2.905 (3.410)	103,182 word features
WANDS metadata-word	480	nDCG@10	0.656	4.284 (5.290)	110,884 word features
WANDS hybrid	480	nDCG@10	0.690	15.866 (18.710)	181,535 word + character features

Table 12 supplies a complete runtime and size account. The combined route feature transformation and inference takes 0.128 ms per AmazonQAC test row and uses 20,006 input features. Context-nearest ranking takes 0.200 ms per row over 6,161 unique training outcomes, while the confidence-gated system takes 0.141 ms per row including route transformation, inference, and candidate selection. WANDS mean retrieval latency rises from 2.905 ms for title-only TF-IDF to 15.866 ms for the hybrid. Figure 8 shows the autocomplete latency-MRR@10 trade-off. These single-process measurements exclude index construction and characterize the implementations tested here, not a production service-level guarantee.

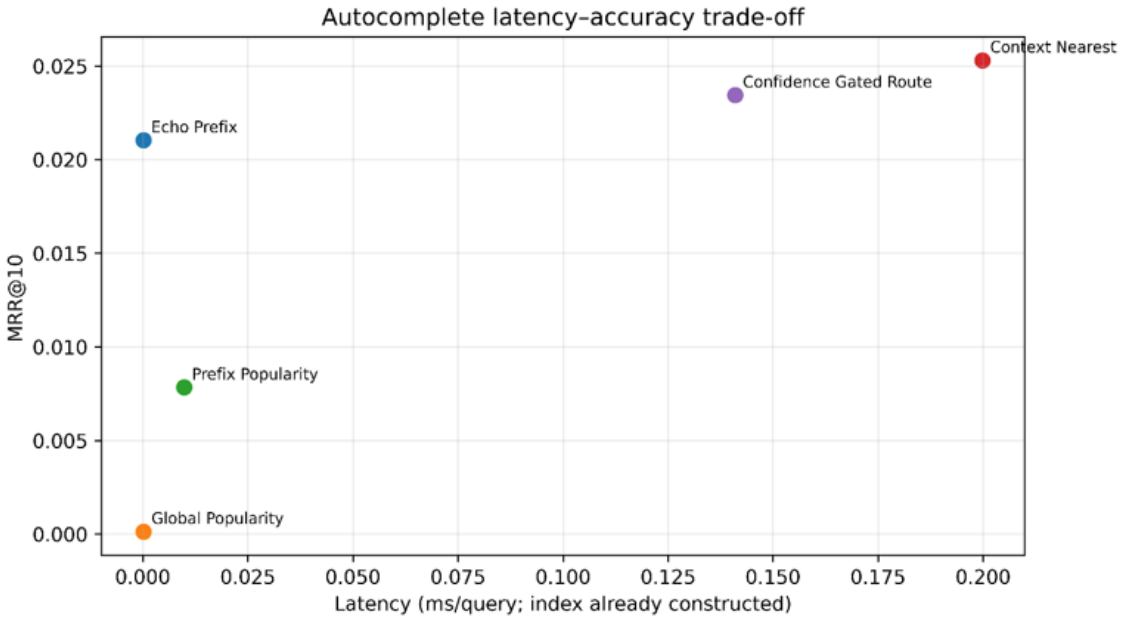

Figure 8. Autocomplete latency-MRR@10 trade-off after index construction

B. Discussion

The main result is that route-aware architecture and exact autocomplete ranking should be judged separately. A router can make entry actions explicit and auditable, yet the confidence-gated method does not outperform context-nearest completion on exact MRR@10 or Success@10. The confidence intervals for their paired differences include zero, and both Holm-adjusted p-values equal 1.000. Consequently, the router is best treated as a conditional control layer rather than as a replacement for the strongest completion ranker.

Route classification itself also warrants qualification. The combined representation improves on the word, character, and heuristic baselines, but five-seed macro-F1 remains 0.330. AUTOSUGGEST is substantially easier than context-aware, direct, fallback, and refinement outcomes. These labels depend partly on the future submitted query, so surface features observed at prefix time cannot fully recover them. A deployment decision would require prospective intent labels, probability calibration, action-specific costs, and a confidence threshold validated against user outcomes.

Candidate coverage explains why all exact QAC scores are modest. Context roughly doubles coverage for 1-4 character prefixes, but the resulting 6.35% ceiling remains low. Echo prefix performs relatively well for long prefixes because some rows are already complete, whereas a training-query index cannot retrieve unseen final strings. A neural suffix generator or catalog-conditioned candidate source could expand coverage, but it should be evaluated separately from ranking so that generation gains are not mistaken for reranking gains.

The WANDS experiment provides the stronger product-search validation because its historical customer queries and human judgments are independent of the target product text (Chen et al., 2022). Richer metadata and character matching significantly improve nDCG@10 and Exact-MRR@10, while Recall@10 and Hit@10 gains remain uncertain and latency increases. This result complements the QAC finding: richer routing or retrieval structure can be useful, but the benefit must be established for the metric and action it is intended to improve.

V. LIMITATIONS

The operational routes are retrospective labels, not independently annotated latent intents, because their definitions use the submitted final query. The study also constructs a custom task within the official AmazonQAC test file; strict chronology and session isolation require excluding more than half of its rows, leaving 951 test cases and small minority-route supports. Candidate generation is limited to prior outcomes and recent terms, and the observed exact-match metrics therefore reflect both coverage and ranking. Source-specific shortcuts are reduced by deriving all route labels within AmazonQAC, but spurious correlations may still arise from time, vocabulary, or rule construction (Wang & Culotta, 2020).

Generalization is bounded by two e-commerce environments. AmazonQAC represents U.S. Amazon searches over a short 2023 period, while WANDS contains 480 Wayfair queries and a furniture-oriented catalog. The Amazon Search Benchmark uses queries generated from target metadata; that design can inflate lexical overlap and is not treated as evidence of real-query retrieval. De-identification reduces direct disclosure risk but does not remove the sensitivity or linkage risk of timestamped behavioral sequences (Amazon, 2024; Pàmies-Estrems & Garcia-Alfaro, 2023). Finally, the latency measurements are hardware- and implementation-specific and omit index construction, concurrency, and network overhead.

VI. CONCLUSION AND RECOMMENDATION

IntentRouter-QAC demonstrates that a lightweight lexical classifier can expose operational search-entry decisions, but it does not establish that routing improves exact autocomplete ranking. Context-nearest completion remains the strongest exact QAC method, and its small advantage over confidence-gated routing is statistically inconclusive. The router's low macro-F1, especially for context-aware and refinement outcomes, further argues for conservative use. A practical architecture should retain context-nearest ranking as the default completion method and invoke route-specific actions only when prospective validation supports them.

Product retrieval should remain a separate, independently judged task. On WANDS, the hybrid lexical representation significantly improves ordering quality over title-only and metadata-word

baselines, although coverage gains are not significant and latency is higher. Future work can add calibrated semantic features, click outcomes, and a compact neural reranker while preserving evidence provenance and explicit fallback reasons, as recommended in evidence-grounded RAG, recommendation-card, and privacy-risk-card research (Jin, 2025b; W. Su et al., 2026; B. Zhang et al., 2025; Zhong, Chen, et al., 2025).

Evaluation should report candidate coverage alongside Success@k and MRR@k, route-level F1 alongside aggregate accuracy, and confidence intervals alongside point estimates. Product retrieval likewise benefits from reporting both ordering metrics and coverage metrics with paired tests. Keeping candidate generation, action routing, completion ranking, and product retrieval distinct makes failure analysis clearer and prevents a strong result in one layer from being presented as evidence for another.

REFERENCES

amazon/AmazonQAC · Datasets at Hugging Face. (n.d.). Retrieved July 31, 2026, from https://huggingface.co/datasets/amazon/AmazonQAC

Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment. *Journal of Technology Informatics and Engineering*, *5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

Bar-Yossef, Z., & Kraus, N. (2011). Context-Sensitive Query Auto-Completion. *Proceedings of the 20th International Conference on World Wide Web*, 107–116. https://doi.org/10.1145/1963405.1963424

Broder, A. (2002). A Taxonomy of Web Search. *ACM SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

Cai, F., & de Rijke, M. (2016). A Survey of Query Auto Completion in Information Retrieval. *Foundations and Trends® in Information Retrieval*, *10*(4), 273–363. https://doi.org/10.1561/1500000055

Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews. *JEECS (Journal of Electrical Engineering and Computer Sciences)*, *11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

Chen, Y., Liu, S., Liu, Z., Sun, W., Baltrunas, L., & Schroeder, B. (2022). *WANDS: Dataset for Product Search Relevance Assessment* (pp. 128–141). https://doi.org/10.1007/978-3-030-99736-6_9

Chen, Y., & Xu, H. (2026). Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access. *International Journal of Graphic Design*, *4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Chenyu Li, Jingwen Bai, & Samuel Wang. (2024). Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM

Applications. *Journal of Advanced Computing Systems*, *4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

Chenyu Li, Wenhao Su, & Eric Zhang. (2023). Lightweight Hallucination Firewall for Enterprise LLM Applications: Evidence Consistency, Self-Checking, and Small-Model Detection on TruthfulQA. *Journal of Advanced Computing Systems*, *3*(1), 49–65. https://doi.org/10.69987/JACS.2023.30104

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/V1/N19-1423

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1383–1392. https://doi.org/10.18653/v1/P18-1128

Everaert, D., Patki, R., Zheng, T., & Potts, C. (2024). AmazonQAC: A Large-Scale, Naturalistic Query Autocomplete Dataset. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1046–1055. https://doi.org/10.18653/v1/2024.emnlp-industry.78

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network*. https://arxiv.org/pdf/1503.02531

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for Natural Language Understanding. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4163–4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372

Jiayi Nie, & David Zheng. (2023). Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal. *Journal of Advanced Computing Systems*, *3*(1), 66–80. https://doi.org/10.69987/JACS.2023.30105

Jiaying Jin, Tina Huang, & Sam Lu. (2024a). A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset. *Journal of Advanced Computing Systems*, *4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

Jiaying Jin, Tina Huang, & Sam Lu. (2024b). Cost-Sensitive Learning, Simulated PU Learning, and One-Class Autoencoding for Extreme-Imbalance Credit Card Fraud Detection. *Journal of Advanced Computing Systems*, *4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

Jin, J. (2025a). Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations. *Journal of Technology Informatics and Engineering*, *4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

Jin, J. (2025b). LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy. *International Journal of Graphic Design*, *3*(2), 397–414. https://doi.org/10.51903/IJGD.V3I2.3698

Jinyi Mu, Yifei Lu, & Michelle Smith. (2023). LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies. *Journal of Advanced Computing Systems*, *3*(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142. https://doi.org/10.1145/775047.775067

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *the Association for Computational Linguistics* (Vol. 2, pp. 427–431). https://aclanthology.org/E17-2068/

Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy. *Journal of Technology Informatics and Engineering*, *4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

Li, C., Liu, G., & Zhao, Z. (2026). Cost-Aware LLM-Style Routing for AIOps Log Analysis: Log Parsing, Anomaly Detection, Fault Diagnosis, and Incident Summarization on LogEval Task Files. *Journal of Technology Informatics and Engineering*, *5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

Pàmies-Estrems, D., & Garcia-Alfaro, J. (2023). On the Self-Adjustment of Privacy Safeguards for Query Log Streams. *Computers & Security*, *134*, 103450. https://doi.org/10.1016/j.cose.2023.103450

Ponte, J. M., & Croft, W. B. (1998). A Language Modeling Approach to Information Retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 275–281. https://doi.org/10.1145/290941.291008

Radlinski, F., & Joachims, T. (2005). Query Chains. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 239–248. https://doi.org/10.1145/1081870.1081899

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. https://doi.org/10.18653/v1/D19-1410

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, *4*(1–2), 1–174. https://doi.org/10.1561/1500000019

Rose, D. E., & Levinson, D. (2004). Understanding User Goals in Web Search. *Proceedings of the 13th International Conference on World Wide Web*, 13–19. https://doi.org/10.1145/988672.988675

Shenghan Lu, & David Zhou. (2023). LLM-Augmented Customer Representation Learning for Next-Purchase Prediction in Online Retail. *Journal of Advanced Computing Systems*, *3*(3), 50–64. https://doi.org/10.69987/JACS.2023.30305

Srinivasan, K., Raman, K., Samanta, A., Liao, L., Bertelli, L., & Bendersky, M. (2022). QUILL: Query Intent with Large Language Models using Retrieval Augmentation and Multi-stage Distillation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 492–501. https://doi.org/10.18653/v1/2022.emnlp-industry.50

Su, W., Chen, S., & Zhao, C. (2025). Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark. *Journal of Technology Informatics and Engineering*, *4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

Su, W., Rao, H., & Ma, E. (2026). Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks. *International Journal of Graphic Design*, *4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

Wang, Z., & Culotta, A. (2020). Identifying Spurious Correlations for Robust Text Classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3431–3440. https://doi.org/10.18653/v1/2020.findings-emnlp.308

Xiaohan Chang, Tong Ye, & Sophia Luo. (2023). LLM-as-Reranker for Personalized Recommendation: Popularity Bias Mitigation and Faithful Natural-Language Explanations on MovieLens 100K. *Journal of Advanced Computing Systems*, *3*(8), 61–78. https://doi.org/10.69987/JACS.2023.30806

Xin, Q. (2026). Self-Supervised Customer Representation Learning for Segmentation and Next-Purchase Prediction on UCI Online Retail. *J-INTECH*, *14*(01), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

Ye, T., Mu, J., & Hunter, J. (2026). Off-Policy Evaluation and Conservative Policy Selection for Slot-Level Dynamic Bidding and Ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering*, *5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

Yifei Lu, Jinyi Mu, & Thao Tran. (2024). Uncertainty-Aware Uplift Modeling for Safer Marketing Targeting: Conformal Prediction and Bayesian Calibration with LCB Policies. *Journal of Advanced Computing Systems*, *4*(5), 84–101. https://doi.org/10.69987/JACS.2024.40507

Yuanzheng Chen, Yitian Zhang, David Chau, & Matt Sherman. (2023). Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset. *Journal of Advanced Computing Systems*, *3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

Yunhe Li. (2024). Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset. *Journal of Advanced Computing Systems*, *4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering*, *4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

Zhong, Z. S., Chenyu Li, & Rao, H. (2026). Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench. *Journal of Technology Informatics and Engineering*, *5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

Ziliang Samuel Zhong, Ruiyan Ma, & Hailey Zhao. (2023). Human-Uncertainty Distillation for Calibrated Vision Models on CIFAR-10H. *Journal of Advanced Computing Systems*, *3*(2), 77–89. https://doi.org/10.69987/JACS.2023.30206