

Original Article

BALANGAY-RAG: Adaptive Pre-Generation Retrieval and Abstention for Filipino-English Code-Switched Queries

John Syte L. Tiopes¹, John Best L. Tiopes², Wilmar S. Orlanda¹, Ronnel A. Desoloc¹

Author Information:

¹Parañaque City College, Coastal Road, cor Victor Medina, Parañaque, 1700 Metro Manila, Philippines

²National University, Mall of Asia Complex, Pasay City, Metro Manila, Philippines

Corresponding Author:

John Syte L. Tiopes
tiopes.johnsyte@paranaquecitycollege.edu.ph

Article History:

Date received: August 13, 2026
Date revised: September 14, 2026
Date accepted: September 24, 2026

Recommended citation:

Tiopes, J.S., Tiopes, J.B., Orlanda, W., & Desoloc, R. (2026). BALANGAY-RAG: Adaptive pre-generation retrieval and abstention for Filipino-English code-switched queries. *Journal of Interdisciplinary Perspectives*, 4(10), 436-447.
<https://doi.org/10.69569/jip.2026.443x>

Abstract. Fixed retrieval depths can waste context or proceed without designated support when language form, evidence availability, and context budgets vary. This controlled pre-generation study evaluates BALANGAY-RAG, combining the BALANGAY-PH benchmark with BALANGAY-RF, a Random Forest controller that selects $k = 1, 2, 4$, or 8 passages or abstains. The reported experiment contains 44 public-information passages, 147 English, Filipino, and Taglish queries, and 1,764 grouped held-out states. BALANGAY-RF achieved 0.80 feasible-evidence retrieval recall, 0.80 gold-passage precision among proceeded states, a 0.11 gold-absent proceed rate, and 15.49 unconditional mean retrieved words. A separately archived post-review reconstruction and reproduction regenerates 1,764 states and 10,584 policy outcomes from 39 official-source URLs, reproducing the qualitative context-coverage trade-off while yielding different numerical operating points. No downstream generator was evaluated.

Keywords: *Abstention; Adaptive retrieval; Code-switching; Retrieval-augmented generation; Taglish.*

Retrieval-augmented generation (RAG) combines a generative model with externally retrieved evidence so responses can be grounded in an inspectable, updatable knowledge collection (Lewis et al., 2020). Before generation begins, however, retrieval already poses a control problem: a system must determine whether supporting evidence is available, how much evidence to retrieve, and whether the retrieved set is sufficient to justify proceeding. Fixed top- k retrieval holds these decisions constant even when query form and resource constraints vary.

Increasing retrieval depth can improve the chance that supporting evidence enters the context, but it also increases downstream context and exposure to irrelevant material. Recent work therefore treats retrieval timing, necessity, and depth as adaptive decisions. FLARE retrieves during generation (Jiang et al., 2023), Self-RAG retrieves on demand and critiques evidence (Asai et al., 2024), Adaptive-RAG routes queries by complexity (Jeong et al., 2024), and RetrievalQA shows that adaptive-retrieval performance can depend strongly on calibration and retrieval-necessity decisions (Zhang et al., 2024). Robustness studies likewise show that irrelevant or noisy context can impair downstream answers (Kim et al., 2024; Yoran et al., 2024). Bandit- and uncertainty-based approaches further demonstrate that retrieval policy is a meaningful optimization target rather than a fixed preprocessing choice (Tang et al., 2025; Moskvoretskii et al., 2025).

Multilingual and code-switched queries add another source of variation. Cross-lingual RAG benchmarks report sensitivity to query and evidence language (Liu et al., 2025), and recent code-switching retrieval work shows that mixed-language queries can remain difficult even for strong multilingual retrievers (Zeng et al., 2026). Tagalog-English code-switching is also documented in Philippine language resources; TweetTaglish contains more than 20,000 Tagalog-English social-media instances (Herrera et al., 2022). These findings motivate a controlled Filipino-English retrieval study and underscore that hand-constructed aligned queries should not be treated as a substitute for naturally occurring Philippine information-seeking language.

Answerability is a related retrieval-stage concern. Component-level RAG evaluation recommends separating retrieval behavior from downstream generation quality (Es et al., 2024), while recent work on unanswerable requests emphasizes rejection alongside answerable performance (Peng et al., 2025). A policy that always retrieves can maximize coverage while still proceeding without designated support; a policy that abstains too often can minimize unsupported proceeding at the cost of usefulness. Retrieval effectiveness, abstention, coverage, and context cost therefore need to be reported together.

This study isolates pre-generation retrieval control. BALANGAY-PH denotes the benchmark corpus and aligned query set; BALANGAY-RF denotes the learned Random Forest controller; and BALANGAY-RAG denotes the integrated retrieval-and-abstention framework. This work does not introduce a new generator, retriever family, or end-to-end RAG architecture. Instead, it combines aligned English-Filipino-Taglish information needs, controlled evidence-availability and context-budget stress conditions, and a learned choice between retrieval depth and abstention. A distinct archival contribution is the versioned post-review reconstruction and reproduction package, which supports rerunning a transparently specified experiment without being represented as the unrecovered historical raw data (Tiopes et al., 2026). All outcome claims remain retrieval-level; we did not evaluate generated-answer accuracy, hallucination, latency, memory, or factual safety after generation.

Table 1 positions BALANGAY-RF against representative adaptive-retrieval approaches. The comparison focuses on decision point, action space, abstention, multilingual scope, and the resource variable being controlled.

Table 1. *Technical Positioning of Adaptive Retrieval Approaches*

Approach	Decision point	Action space	Abstention	Multilingual focus	Resource measure
Adaptive-RAG (Jeong et al., 2024)	Pre-generation routing	Retrieval strategy by complexity	Not central	Not central	Strategy/complexity
Uncertainty-adaptive retrieval (Moskvoretskii et al., 2025)	Pre-generation	Retrieve/adapt using uncertainty	Policy-dependent	Not central	Retrieval need/uncertainty
MBA-RAG (Tang et al., 2025)	Pre-generation	Bandit-selected retrieval action	Not central	Not central	Adaptive action cost/reward
BALANGAY-RF (this study)	Pre-generation	$k = 1, 2, 4, 8$, or abstain	Explicit	English/Filipino/Taglish	Retrieved-word budget and evidence availability

Accordingly, the study addressed three questions: (1) whether reciprocal-rank retrieval effectiveness differs among aligned English, Filipino, and Taglish queries when the underlying information need is held constant; (2) what efficiency-coverage-abstention trade-off is created by learned adaptive control relative to fixed and confidence-gated retrieval; and (3) how that trade-off changes as evidence availability and context budgets are reduced. The purpose is to characterize a reproducible operating point for resource-aware retrieval control in a

controlled Filipino-English setting.

Methodology

Research Design

The study used a controlled computational proof-of-concept design with no human participants. It isolates the pre-generation retrieval-control layer, so retrieval ranking, evidence availability, context capacity, and abstention can be evaluated without confounding from a downstream language model. In this manuscript, “gold-absent proceeding” refers only to a retrieval-level proxy: proceeding without the designated benchmark passage. It should not be interpreted as generated-answer error or demonstrated real-world harm.

Benchmark and Query Construction

BALANGAY-PH contains 44 concise evidence passages derived from publicly accessible information issued by six Philippine public agencies: the Philippine Atmospheric, Geophysical and Astronomical Services Administration, the Philippine Statistics Authority, the Department of Social Welfare and Development, the Commission on Higher Education, the Department of Health, and the Department of Information and Communications Technology. The documented construction rule retained attributable public-information material that could be represented as a short, self-contained evidence unit for one identifiable information need. The corpus spans disaster information, civil registration, social protection, higher education, public health, and digital government services. Public information was paraphrased for benchmarking while source provenance was retained, and the 44 passages were assigned to four evidence shards of 11 passages each.

The surviving materials for the reported experiment do not contain the original per-agency passage counts, passage-length distribution, per-source access dates, granular paraphrasing log, author-by-author query-construction record, exact normalization-lexicon history, or original passage-to-shard mapping. Because these are historical provenance fields, this revision does not reconstruct them as if they were original. Their absence remains a reproducibility limitation. A separate, versioned post-review reconstruction and reproduction is publicly archived for transparency and rerunability; it is explicitly distinct from the unrecovered original artifacts (Tiopes et al., 2026).

Each evidence item generated three aligned answerable queries in English, Filipino, and Filipino-English code-switched Taglish, yielding 132 answerable queries. Five additional out-of-database information-need families were represented in the same three language forms, yielding 15 unanswerable queries and 147 total queries. Taglish was operationalized as a controlled query containing both Filipino and English lexical material for the same information need. No independent native-speaker naturalness panel documented the original query set, so we treat these forms as controlled code-switched test variants rather than validated naturalistic Taglish. All variants of one information need shared a group identifier, and grouped five-fold cross-validation kept all language, availability, budget, and candidate rows from that information need in the same fold.

Gold-Evidence Definition and Validation Scope

Every answerable query in the reported experiment was linked to one designated gold passage. The original metrics therefore measure recovery of that designated passage, not the broader proposition that every non-gold passage is irrelevant. This matters because a different passage may overlap with, partially support, update, or contradict the designated item. The original record does not document independent second-pass relevance adjudication, multiple-gold labels, or a dedicated hard-negative challenge set. Accordingly, the reported policy metrics remain single-gold metrics and are interpreted as benchmark-passage recovery rather than exhaustive factual relevance.

Archived Post-Review Reconstruction and Reproduction

Following Stage 3 review, the authors archived BALANGAY-RAG Post-Review Reconstruction and Independent Reproduction Package, Version 1.0.0 (Tiopes et al., 2026). The archive contains 44 reconstructed evidence passages grounded in 39 unique official Philippine government-source URLs, 147 aligned query variants, source and paraphrase metadata, an explicit 42-entry reproduction lexicon, a deterministic four-shard reproduction assignment, executable code, 1,764 newly generated resource states, 3,571 budget-feasible candidate-level held-out Random Forest predictions, and 10,584 newly generated policy outcomes. The archive does not include the

unrecovered original dataset or code.

The reproduction preserves the documented word and character TF-IDF settings, the 0.60/0.40 hybrid weighting, candidate depths $k = 1, 2, 4$, and 8 , the 60-, 120-, and 240-word budgets, four evidence-availability levels, grouped five-fold cross-validation, and the reported Random Forest hyperparameters. We newly fix unrecovered historical implementation choices and explicitly label them as reproduction specifications. The reconstructed evidence units range from 23 to 50 words ($M = 39.52$, median = 40.50, $SD = 6.60$). Under the documented reproduction rules, 951 states are feasible, and 3,571 candidate rows are budget-feasible, comprising 1,906 positive and 1,665 negative candidate labels.

We use the archived reproduction as a reproducibility and sensitivity check rather than a replacement for the reported raw experiment. Because the historical and reconstructed evidence sets are not identical, exact numerical agreement is not expected. This release does not include independent human relevance adjudication; consequently, the single-gold interpretation remains in force, and we make no multiple-gold or exhaustive factual-relevance claim. Table 2 summarizes the reported benchmark design.

Table 2. *BALANGAY-PH Benchmark and Experimental Design*

Factor	Setting
Evidence corpus	44 passages; 6 public agencies
Queries	147 total; 132 answerable + 15 unanswerable
Languages	English, Filipino, Taglish
Availability	Stable 4/4; Degraded 3/4; Intermittent 2/4; Local-only 1/4 shards
Context budgets	60, 120, 240 retrieved words
Candidate k	1, 2, 4, 8; or abstain
Held-out states	$147 \times 4 \times 3 = 1,764$
Policy outcomes	10,584

Hybrid Retrieval Layer and Resource-Stress Conditions

The retrieval layer used transparent term frequency-inverse document frequency (TF-IDF) representations (Manning et al., 2008). Word features used unigrams and bigrams, character features used within-word 3- to 5-grams, and both used sublinear term frequency. A small normalization lexicon expanded common Filipino interrogatives and selected domain expressions with English equivalents before retrieval. The original record does not contain the historical lexicon’s exact entry count, complete construction history, or authorship. The archived reproduction separately supplies an explicit 42-entry lexicon as a reproduction specification rather than a recovered historical artifact. The BALANGAY hybrid score combined 0.60 of the normalized-query word score and 0.40 of the normalized-query character score; these weights were fixed for the reported experiment. Because character TF-IDF equaled or exceeded the hybrid on several language metrics, we treat the hybrid as a transparent design choice rather than a universally superior retriever.

We operationalized resource stress through evidence availability and retrieved-word budget. Stable exposed 4/4 shards, Degraded 3/4, Intermittent 2/4, and Local-only 1/4. These are evidence-availability proxies, not simulations of bandwidth, latency, or packet loss. The reported experiment used one fixed shard subset per availability level, so availability is partly confounded with the passages present in those subsets; the exact historical passage-to-shard mapping is not present in the surviving original files. The archived reproduction therefore uses a separately labeled deterministic balanced shard assignment. Context budgets of 60, 120, and 240 retrieved words constrained each candidate action; a candidate k was considered budget-feasible only if its complete selected passage set fit within the budget. Tokenizer-specific truncation was not evaluated. Candidate depths were $k = 1, 2, 4$, and 8 , producing $147 \times 4 \times 3 = 1,764$ held-out resource states.

BALANGAY-RF Controller

For each resource state, the controller used query length, language indicators, evidence-availability indicators, budget, the top retrieval scores, score margin, normalized entropy over the top eight scores, agreement between word and character retrievers, available-corpus fraction, and candidate-specific context features. A candidate-level Random Forest classifier (Breiman, 2001) used 160 trees, maximum depth 14, minimum leaf size 2, balanced-subsample class weighting, and deterministic fold-specific seeds. Local-only still exposes 11 passages, so eight retrieval scores are available without padding. For every budget-feasible k , the classifier estimated whether the

designated gold passage would be covered; BALANGAY-RF selected the smallest k with predicted probability at least 0.50 and otherwise abstained. The 0.50 value is an operating threshold, not evidence of probability calibration. The reported historical experiment did not include Brier scores, reliability curves, or threshold-sweep risk-coverage analyses. Hyperparameters and the threshold were fixed before held-out scoring, but the surviving original record does not identify a separate hyperparameter-development set or exact candidate-level class counts. The archived reproduction supplies the full feature implementation, held-out candidate probabilities, and candidate labels needed for a separate calibration check.

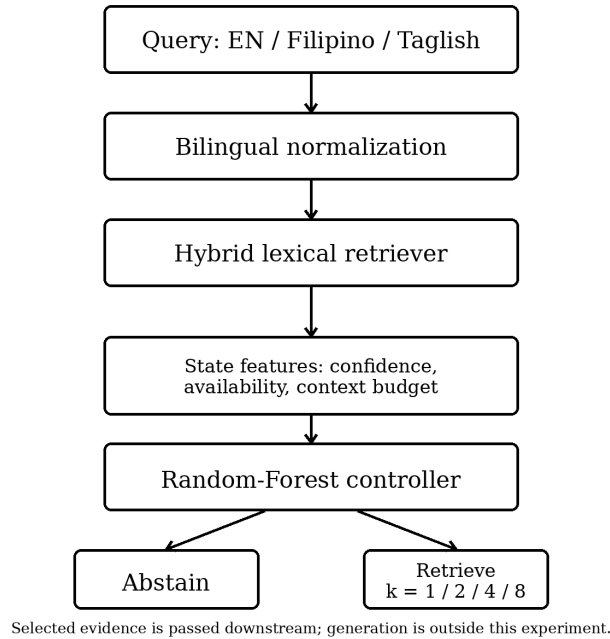


Figure 1. BALANGAY-RAG Pre-Generation Retrieval and Abstention Controller

Data Gathering Procedure

Data generation proceeded in four stages. First, we ran each aligned query against the shards exposed under the current availability condition. Second, the hybrid retriever produced ranked evidence and candidate-specific context features under the current word budget. Third, grouped cross-validation generated held-out BALANGAY-RF decisions and baseline outcomes. Fourth, the selected action, retrieved context, designated-gold coverage, abstention status, and resource-state identifiers were recorded. The reported design yielded 1,764 held-out states and 10,584 policy-state outcomes.

Baselines and Evaluation Metrics

We evaluated fixed $k = 1, 2, 4$, and 8 policies, a confidence-gated $k = 4$ baseline, and BALANGAY-RF on the same held-out states. The confidence gate selected a top-score threshold using training-fold answerability $F1$; it abstained when the held-out top score fell below that threshold. Every selected k remained subject to the current word budget. Primary retrieval-level outcomes were feasible-evidence retrieval recall, gold-passage precision among proceeded states, coverage, gold-absent proceed rate, over-abstention rate, and mean retrieved context. These names intentionally avoid implying generated-answer accuracy or demonstrated system harm.

A state was feasible when the designated gold passage was available, and at least one candidate k could include it within the word budget. The oracle target was the smallest feasible k containing that passage; inaccessible evidence, unanswerable queries, and states with no budget-feasible supporting set received a target of 0. Let F denote feasible states, P denote proceeded states, S denote states whose selected evidence contains the designated gold passage, and N denote all states. Feasible-evidence retrieval recall is $|S \cap F| / |F|$; gold-passage precision among proceeded states is $|S \cap P| / |P|$; and coverage is $|P| / N$. Gold-absent proceeding is the proportion of all

states that proceed without the designated gold passage, while over-abstention is the proportion that abstain despite feasibility. Unconditional mean context averages retrieved words over all states, with abstentions assigned zero. A consistency audit of the reported matched-state counts recovers $N = 1,764$ and a common feasible denominator $F = 949$; these reconstructed denominators are labeled as post-review recoveries rather than original raw-output fields.

Data Analysis Procedure

For language effects, aligned full-availability reciprocal ranks from the hybrid retriever were compared across the same 44 answerable information-need blocks using a Friedman test, with Kendall's W as the omnibus effect size. Paired Wilcoxon tests with Holm adjustment were used for post-hoc language comparisons. Policy comparisons used matched held-out states. Cluster-bootstrap confidence intervals resampled the underlying information-need group so language, availability, and budget variants from one information need remained dependent. Binary matched-state comparisons used a two-sided exact binomial test on discordant pairs, equivalent to exact McNemar testing. We used Wilcoxon signed-rank tests for paired context comparisons. We interpreted these state-level tests descriptively when cluster-adjusted equivalents were unavailable, because treating all 1,764 states as independent can create pseudoreplication. Numerical results are rounded to two decimals and p -values to three decimals or reported as $p < .001$.

Ethical Considerations and Reproducibility

The study collected no human-participant or personally identifying research data. The project nevertheless underwent institutional review and received ethical clearance from the Parañaque City College Center for Research and Publication Ethics Review Board on July 22, 2026. The authors paraphrased government-source information for benchmarking, and retained source titles and official URLs in the archived reconstruction. Treat the benchmark as a dated research snapshot rather than a substitute for current official guidance, and check reuse obligations against the source-specific terms of the originating agencies. Reproducibility is supported by grouped cross-validation, fixed held-out thresholds, deterministic seeds, matched-state comparisons, and the public Version 1.0.0 reconstruction and reproduction package archived at <https://doi.org/10.5281/zenodo.22274900> (Tiopes et al., 2026). The archive is explicitly distinguished from the unrecovered historical experimental artifacts.

Results and Discussion

Static Retrieval Effectiveness by Language

Table 3. Full-Availability Retrieval by Query Language (MRR / Recall@4)

Retriever	English	Filipino	Taglish
Word TF-IDF	0.87 / 0.93	0.86 / 0.93	0.88 / 0.91
Char TF-IDF	0.90 / 1.00	0.87 / 0.93	0.90 / 0.98
BALANGAY hybrid	0.89 / 0.96	0.87 / 0.93	0.89 / 0.96

Under full evidence availability, word TF-IDF achieved mean reciprocal rank (MRR) values of 0.87, 0.86, and 0.88 for English, Filipino, and Taglish, respectively; character TF-IDF achieved 0.90, 0.87, and 0.90; and the BALANGAY hybrid achieved 0.89, 0.87, and 0.89. Hybrid Recall@4 was 0.96, 0.93, and 0.96. Across 44 complete answerable information-need blocks, the aligned hybrid ranks did not differ significantly by language, Friedman $\chi^2(2) = 4.24$, $p = .120$, with Kendall's $W = 0.05$. The reported Holm-adjusted paired comparisons were also non-significant, although the individual adjusted p -values were not retained in the surviving original record. Because no equivalence margin or equivalence test was prespecified, these findings indicate a lack of detected difference on this benchmark, not proof of language equivalence. Table 3 reports the language-specific retrieval metrics, and Figure 2 visualizes the hybrid MRR values.

The controlled design likely contributes to the similar language scores: each information need is aligned across three language forms, the corpus is small, and explicit normalization reduces uncontrolled lexical variation. This differs from broader cross-lingual and code-switching benchmarks that report language-sensitive retrieval behavior (Liu et al., 2025; Zeng et al., 2026). A stronger external-validity test would use naturally authored Taglish, varied switch points, independent paraphrases, and denser corpora in which lexical overlap is less informative.

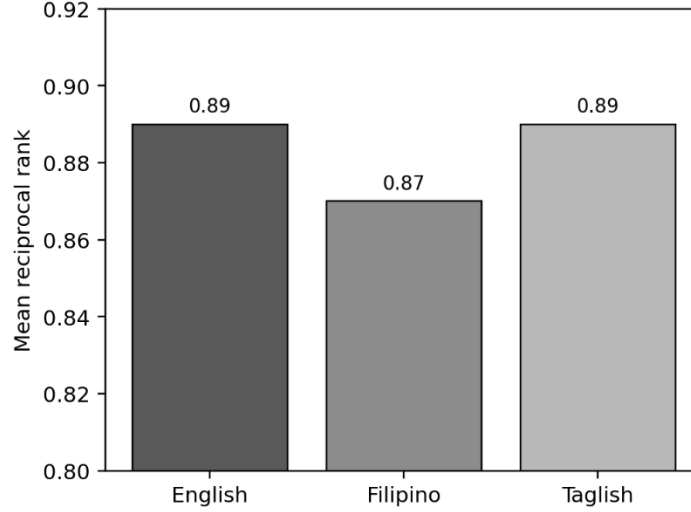


Figure 2. Hybrid Retriever Mean Reciprocal Rank for Aligned English, Filipino, and Taglish Queries

Overall Retrieval-Policy Trade-off

Table 4. Selected Overall Retrieval-Policy Performance

Policy	F-Rec	Gold-Prec	Coverage	Gold-absent	Over-Abst	P / A	Context U / C
Fixed-1	0.88	0.47	1.00	0.53	0.00	1,764 / 0	21.69 / 21.69
Fixed-2†	≈0.95	≈0.51	1.00	≈0.49	0.00	1,764 / 0	≈44.15 / ≈44.15
Fixed-4	0.99	0.53	1.00	0.47	0.00	1,764 / 0	74.27 / 74.27
Fixed-8	1.00	0.54	1.00	0.46	0.00	1,764 / 0	104.78 / 104.78
Confidence-4	0.91	≈0.81	0.61	0.12	NR	1,074 / 690	44.50 / ≈73.09
BALANGAY-RF	0.80	0.80	0.54	0.11	NR	957 / 807	15.49 / ≈28.55

Note. F-Rec = feasible-evidence retrieval recall; Gold-Prec = gold-passage precision among proceeded states; Gold-absent = proportion of all states proceeding without the designated gold passage; Over-Abst = over-abstention rate; P / A = proceed / abstain counts; Context U / C = unconditional / proceeding-conditional mean retrieved words. "NR" indicates not recoverable from the submitted aggregate results. †Fixed-2 values are approximate figure-derived values and are excluded from new inferential claims.

Across the 1,764 held-out states, a post-review consistency audit recovers 949 feasible states. Structurally, the design contains 180 unanswerable states, 594 answerable states whose designated gold passage is inaccessible under the current shard condition, and 990 accessible-answerable states; 41 of the latter do not satisfy the final budget-feasibility rule. Fixed policies always proceed, whereas the two selective policies abstain in some states. Reconstructed overall proceed / abstain counts are 1,074 / 690 for Confidence-4 and 957 / 807 for BALANGAY-RF. These counts are reported as consistency recoveries from the published aggregate relationships, not as replacements for the missing original state-level table. Table 4 summarizes the selected policy metrics, and Figure 3 plots feasible-evidence recall against unconditional mean retrieved context.

Relative to Fixed-4, BALANGAY-RF reduced unconditional mean context by 79.10% and gold-absent proceeding by 35.66 percentage points. Cluster-bootstrap differences were -58.81 words (95% CI [-62.06, -55.52]) for context, -0.36 (95% CI [-0.45, -0.27]) for gold-absent proceeding, and -0.19 (95% CI [-0.29, -0.09]) for feasible-evidence recall. The result is therefore an operating-point shift rather than universal superiority: BALANGAY-RF is substantially more context-conservative, while fixed and confidence-gated policies preserve more coverage. Because unconditional context assigns zero words to abstentions, part of the apparent saving is mechanically caused by abstention. The post-review overall proceeding-conditional means are approximately 28.55 words for BALANGAY-RF and 73.09 for Confidence-4; condition-specific proceeding-conditional means require the original policy rows.

This trade-off is consistent with adaptive-retrieval work that treats retrieval need and depth as policy decisions and with robustness studies showing that excess context can be distracting (Jiang et al., 2023; Asai et al., 2024; Jeong et al., 2024; Kim et al., 2024; Yoran et al., 2024; Zhang et al., 2024). It also illustrates why coverage and

abstention must accompany context-efficiency claims. Confidence-4 used substantially more context than BALANGAY-RF but retained higher feasible recall with a similar gold-absent proceed rate. Transparent heuristic baselines therefore remain important comparators when evaluating learned retrieval control.

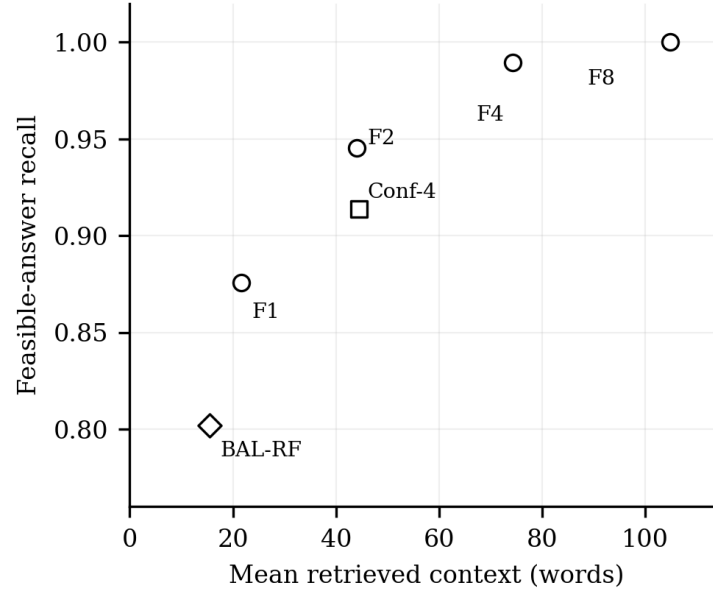


Figure 3. Feasible-evidence retrieval recall versus unconditional mean retrieved context. F1, F2, F4, and F8 denote fixed $k = 1, 2, 4$, and 8 policies; Conf-4 denotes confidence-gated $k = 4$; BAL-RF denotes BALANGAY-RF.

Evidence Availability and Context Budget

Table 5. Adaptive Policies by Evidence Availability (F-Rec / Gold-Prec / Context)

State	Confidence-4	BALANGAY-RF
Stable	0.93 / 0.91 / 61.39	0.91 / 0.91 / 24.14
Degraded	0.91 / 0.82 / 52.03	0.82 / 0.78 / 20.33
Intermittent	0.89 / 0.78 / 37.98	0.70 / 0.72 / 11.66
Local-only	0.94 / 0.57 / 26.62	0.55 / 0.55 / 5.83

Evidence availability changed both adaptive policies. Before the final budget-feasibility test, each availability-by-budget slice contains 147 states: Stable has 132 accessible-answerable and 15 unanswerable states; Degraded has 99 accessible-answerable, 33 inaccessible-answerable, and 15 unanswerable; Intermittent has 66, 66, and 15; and Local-only has 33, 99, and 15. Across the three budgets, these totals become 396/0/45, 297/99/45, 198/198/45, and 99/297/45, respectively. The 41 accessible-answerable states that fail the final feasibility rule cannot be assigned to specific availability-budget cells without the original state-level outputs. Within the reported aggregates, BALANGAY-RF becomes increasingly conservative as availability decreases: feasible-evidence recall falls from 0.91 in Stable to 0.55 in Local-only, while unconditional context falls from 24.14 to 5.83 words. Confidence-4 preserves higher recall but uses more context. Cross-condition recall comparisons must therefore be interpreted with the changing feasible-state denominator in mind. Table 5 summarizes the availability-specific results, and Figure 4 shows gold-passage precision among proceeded states across availability conditions.

Context budget changed resource use more than it changed the qualitative policy ordering. BALANGAY-RF used 11.65, 16.54, and 18.28 unconditional mean words at 60-, 120-, and 240-word budgets, compared with 25.89, 53.81, and 53.81 for Confidence-4. BALANGAY-RF therefore remained selective even when it could admit more retrieved words. Because budgets are measured in words rather than model-specific tokens, these results should not be translated directly into token, latency, memory, or energy savings. Table 6 summarizes performance across the three word-budget conditions.

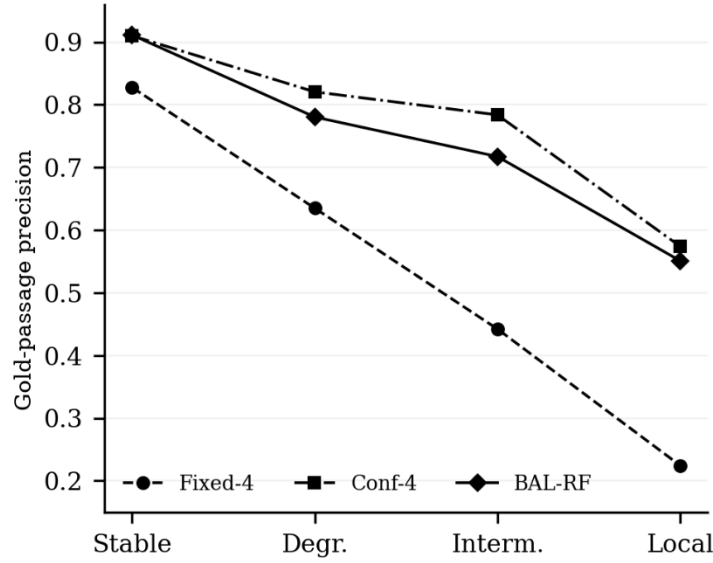


Figure 4. Gold-passage precision among proceeded states as evidence availability decreases. The vertical axis is conditional on proceeding: $|S \cap P|/|P|$, where S denotes selections containing the designated gold passage, and P denotes proceeded states.

Table 6. Adaptive Policies by Context Budget (F-Rec / Gold-Prec / Context)

Budget	Confidence-4	BALANGAY-RF
60	0.93 / 0.78 / 25.89	0.80 / 0.81 / 11.65
120	0.92 / 0.82 / 53.81	0.81 / 0.79 / 16.54
240	0.89 / 0.82 / 53.81	0.80 / 0.79 / 18.28

For a retrieval-stage deployment, the choice of policy depends on the operating objective. BALANGAY-RF is attractive when minimizing passed context is a primary constraint, whereas Confidence-4 is preferable when preserving feasible-evidence recall and coverage matters more. These results do not establish that either policy is sufficient for production use; the benchmark is small and no downstream generator was evaluated.

Language-Specific Adaptive Behavior and Matched Tests

Table 7. BALANGAY-RF Performance by Query Language

Language	F-Rec	Gold-Prec	Gold-absent	Context
English	0.80	0.77	0.13	15.85
Filipino	0.80	0.80	0.11	15.96
Taglish	0.81	0.82	0.10	14.66

Across language subsets, BALANGAY-RF feasible-evidence recall was 0.80 for English, 0.80 for Filipino, and 0.81 for Taglish; gold-passage precision among the preceding states was 0.77, 0.80, and 0.82; and unconditional mean context was 15.85, 15.96, and 14.66 words. Confidence-4 preserved higher feasible recall in each language (0.94, 0.91, and 0.89) but used 46.80, 45.72, and 40.99 words. These descriptive results do not establish equivalence or the absence of a language-specific failure mode; they only show that no large separation was detected on this controlled benchmark. Table 7 summarizes BALANGAY-RF performance by query language.

Matched-state comparisons confirm the main trade-offs. Against Confidence-4, BALANGAY-RF succeeded on 7 feasible states where Confidence-4 failed, while Confidence-4 succeeded on 113 where BALANGAY-RF failed; a two-sided exact binomial test on the 120 discordant feasible states, equivalent to exact McNemar testing, gave $p < .001$. BALANGAY-RF also used less context (reported Wilcoxon signed-rank $p < .001$); the original Wilcoxon statistic and rank-based effect size were not retained in the surviving record and therefore cannot be reconstructed. For gold-absent proceeding, 42 states favored BALANGAY-RF and 31 favored Confidence-4, yielding $p = .242$ in the state-level exact comparison. Against Fixed-4, BALANGAY-RF avoided gold-absent proceeding on 639

discordant states versus 10 in the opposite direction ($p < .001$), while Fixed-4 succeeded on 178 feasible states where BALANGAY-RF failed ($p < .001$). Because many states share the same underlying information need, these state-level p -values are descriptive; cluster-resampled confidence intervals are the preferred uncertainty summary when available.

Post-Review Reproduction Check

Table 8. *Reported Original and Archived Reproduction Operating Points*

Policy	Orig. F-Rec	Reprod. F-Rec	Orig. Gold-abs.	Reprod. Gold-abs.	Orig. Context	Reprod. Context
Fixed-1	0.88	0.953	0.53	0.486	21.69	38.72
Fixed-2†	≈0.95	0.991	≈0.49	0.466	≈44.15	65.17
Fixed-4	0.99	1.000	0.47	0.461	74.27	91.27
Fixed-8	1.00	1.000	0.46	0.461	104.78	91.27
Confidence-4	0.91	0.957	0.12	0.081	44.50	54.93
BALANGAY-RF	0.80	0.875	0.11	0.103	15.49	22.69

Note. Original values are the reported or post-review recovered aggregates in Table 4; Fixed-2 original values are approximate. Reproduction values are newly generated from the archived Version 1.0.0 reconstruction and do not replace the historical results.

The public archived reproduction provides a direct rerunability check on the reconstructed benchmark. It yields 951 feasible states, compared with the 949-state denominator recovered from the reported aggregate relationships. Under the reproduction, BALANGAY-RF achieves 0.875 feasible-evidence recall, 0.821 gold-passage precision among proceeded states, 0.574 coverage, a 0.103 gold-absent proceed rate, 22.69 unconditional mean retrieved words, and 39.51 words conditional on proceeding. Confidence-4 achieves 0.957 feasible-evidence recall, 0.864 gold-passage precision, 0.597 coverage, a 0.081 gold-absent proceed rate, 54.93 unconditional mean words, and 92.01 words conditional on proceeding. Thus, the qualitative operating-point conclusion is reproduced — BALANGAY-RF remains more context-conservative while Confidence-4 preserves more feasible-evidence recall — but the numerical operating points are not identical.

The archived candidate-level reproduction contains 3,571 held-out probability/label rows (1,906 positive and 1,665 negative) and yields an overall candidate-level Brier score of 0.093. This calibration result pertains only to the post-review reconstruction and is not retroactively attributed to the original experiment. The reconstructed evidence units are longer than the historical context totals imply; under the reproduction’s complete-passage budget rule, $k = 8$ is never budget-feasible, and Fixed-8 therefore falls back to the same largest eligible action as Fixed-4. This is a property of the reconstructed rerun, not evidence about the unrecovered historical Fixed-8 implementation. The archived package includes exact state-level denominators, proceeding/abstention flags, candidate probabilities, and policy outputs.

Limitations

The study has six main limitations. First, the reported corpus is small, and the surviving submission materials do not include the original per-item provenance metadata or raw experimental artifacts; the archived reconstruction improves transparency but does not recover historical provenance. Second, the query set uses controlled aligned English, Filipino, and Taglish forms, includes only five underlying unanswerable families, and has no documented independent native-speaker naturalness validation. Third, relevance in the reported experiment is single-gold; the present release does not include independent human relevance adjudication, multiple-gold labels, or a dedicated hard-negative challenge set. Fourth, the reported availability conditions use fixed shard subsets, the historical shard mapping is unavailable, and context is measured in words rather than tokenizer-specific tokens; the archived deterministic shard assignment is a new reproduction specification. Fifth, we evaluate only one learned-controller family. The original Random Forest probabilities were not calibrated, while the post-review Brier score of 0.093 applies only to the reconstructed rerun and does not calibrate the historical experiment. Sixth, state-level dependence can create pseudoreplication, and several condition-specific historical statistics remain unrecoverable even though the archived reproduction now provides complete new state- and policy-level outputs. No downstream generator was executed.

These limitations constrain the study’s claims. Grouped cross-validation, deterministic seeds, matched comparisons, and cluster-resampled intervals support internal consistency, while the public reconstruction and reproduction package materially improves rerunability and auditability. However, construct validity remains tied

to designated-gold recovery and word-count context, and external validity remains limited by the small public-information corpus and controlled language variants. The archived package should therefore be interpreted as a transparent reproduction and sensitivity analysis, not as recovery of the historical raw experiment. Independent human relevance validation would still be required for any future claim of independently adjudicated relevance.

Conclusion

BALANGAY-RAG demonstrates a controlled way to treat retrieval depth and abstention as pre-generation resource decisions for Filipino-English code-switched queries. In the reported experiment, BALANGAY-RF substantially reduced the amount of evidence passed downstream and reduced gold-absent proceeding relative to unconditional fixed retrieval, but it also sacrificed feasible-evidence recall and coverage. Confidence gating remained competitive when we prioritized evidence coverage. The archived post-review reproduction yields different numerical operating points but reproduces the same qualitative context-versus-coverage trade-off, reinforcing that the central contribution is an explicit retrieval-stage framework for examining how context economy, evidence coverage, and abstention move together under changing resource conditions rather than a universally superior controller.

The results support retrieval-stage claims only. Future work should expand naturally authored Taglish and difficult in-domain unanswerables, obtain independent human relevance judgments with multiple acceptable evidence passages, randomize evidence-shard stress conditions, evaluate character-only and no-normalization controller ablations, extend probability-calibration and risk-coverage analyses, compare additional controller families, and connect the same decisions to downstream generators. End-to-end evaluation is necessary before claiming improvements in answer faithfulness, hallucination reduction, latency, memory, or energy.

Contributions of Authors

John Syte L. Tiopes: Conceptualization, methodology, formal analysis, data curation, visualization, writing - original draft, and project administration.

John Best L. Tiopes: Investigation, data curation, validation, and writing - review and editing.

Wilmar S. Orlandia: Supervision, methodology, validation, and writing - review and editing.

Ronnel A. Desoloc: Supervision, validation, and writing - review and editing.

All authors reviewed and approved the final manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interests

The authors declare no conflicts of interest.

Acknowledgment

The authors have no acknowledgments to declare.

References

- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. The Twelfth International Conference on Learning Representations. <https://openreview.net/forum?id=hSyW5go0v8>
- Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Herrera, M., Aich, A., & Parde, N. (2022). TweetTaglish: A dataset for investigating Tagalog-English code-switching. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2090–2097. <https://aclanthology.org/2022.lrec-1.225/>
- Jeong, S., Baek, J., Cho, S., Hwang, S.J., & Park, J. (2024). Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 7036–7050. <https://doi.org/10.18653/v1/2024.naacl-long.389>
- Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 7969–7992. <https://doi.org/10.18653/v1/2023.emnlp-main.495>
- Kim, Y., Kim, H.J., Park, C., Park, C., Cho, H., Kim, J., Yoo, K. M., Lee, S.-G., & Kim, T. (2024). Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. Findings of the Association for Computational Linguistics: EMNLP 2024, 2421–2431. <https://doi.org/10.18653/v1/2024.findings-emnlp.136>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
- Liu, W., Trenous, S., Ribeiro, L.F.R., Byrne, B., & Hieber, F. (2025). XRAG: Cross-lingual retrieval-augmented generation. Findings of the Association for Computational Linguistics: EMNLP 2025, 15669–15690. <https://doi.org/10.18653/v1/2025.findings-emnlp.849>
- Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
- Moskvoretskii, V., Marina, M., Salnikov, M., Ivanov, N., Pletenev, S., Galimzianova, D., Krayko, N., Kononov, V., Nikishina, I., & Panchenko, A. (2025). Adaptive retrieval without self-knowledge? Bringing uncertainty back home. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6355–6384. <https://doi.org/10.18653/v1/2025.acl-long.319>
- Peng, X., Choubey, P.K., Xiong, C., & Wu, C.-S. (2025). Unanswerability evaluation for retrieval augmented generation. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 8452–8472. <https://doi.org/10.18653/v1/2025.acl-long.415>
- Tang, X., Gao, Q., Li, J., Du, N., Li, Q., & Xie, S. (2025). MBA-RAG: A bandit approach for adaptive retrieval-augmented generation through question complexity. Proceedings of the 31st International Conference on Computational Linguistics, 3248–3254. <https://aclanthology.org/2025.coling-main.218/>

- Tiopes, J.S., Tiopes, J.B., Orlanda, W., & Desoloc, R. (2026). BALANGAY-RAG Post-Review Reconstruction and Independent Reproduction Package [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.22274900>
- Yoran, O., Wolfson, T., Ram, O., & Berant, J. (2024). Making retrieval-augmented language models robust to irrelevant context. The Twelfth International Conference on Learning Representations. <https://openreview.net/forum?id=ZS4m74kZpH>
- Zeng, Q., Lu, Y., Zhou, Z., Qi, H., Yu, P., Zhao, F., Yanaka, H., Xuan, W., & Yokoya, N. (2026). Code-switching information retrieval: Benchmarks, analysis, and the limits of current retrievers. Findings of the Association for Computational Linguistics: ACL 2026, 13055–13071. <https://doi.org/10.18653/v1/2026.findings-acl.636>
- Zhang, Z., Fang, M., & Chen, L. (2024). RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. Findings of the Association for Computational Linguistics: ACL 2024, 6963–6975. <https://doi.org/10.18653/v1/2024.findings-acl.415>