JOMLAI: Journal of Machine Learning and Artificial Intelligence
Vol. 5, No. 3, September 2026, pp. 142~152
ISSN: 2828-9102 (Print) | 2828-9099 (Online)
DOI: 10.55123/jomlai.v5i3.9547

Confidence-Gated Semantic Retrieval for Indonesian Customer-Service Intent Classification

Marco Julius Andreas Bakkara[1], Masherlina[2]
[1,2]Teknologi Rekayasa Instrumentasi, Fakultas Vokasi, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received 02 Juli 2026
Revised 24 Agustus 2026
Accepted 29 September 2026

Keywords:

Abstention Mechanism
Confidence Calibration
Indonesian Language
Intent Classification
Sentence Embedding

ABSTRACT

Customer-service chatbots deployed in low-resource languages must decide not only which intent a message expresses but also when to decline answering. This study evaluated a confidence-gated retrieval architecture for Indonesian intent classification in which a deterministic keyword layer, a multilingual sentence encoder, and an explicit abstention rule were combined. A corpus of twenty-two intents containing three thousand nine hundred sixty training patterns and one thousand fifty-six held-out test patterns was constructed, covering commerce enquiries, complaint handling, and instrumentation engineering questions. Two retrieval backends were compared under identical decision logic: a lexical backend that fused word and character term weighting with a naive Bayes component, and a semantic backend built on multilingual sentence embeddings. The semantic backend reached an accuracy of 82.67 percent and a macro-averaged F-measure of 83.04 percent, exceeding the lexical backend by 11.74 and 12.31 percentage points, at a mean response latency of 30.66 milliseconds on a single processor. Raising the confidence threshold increased selective accuracy from 71.06 to 82.82 percent while coverage fell to 61.17 percent, which quantified the price of caution. Of 100 balanced out-of-scope probes, 93 were rejected (93.00 percent). The deterministic layer resolved just over half of the queries and carried the higher accuracy of the two decision paths.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Masherlina,
Teknologi Rekayasa Instrumentasi,
Universitas Sumatera Utara,
Email: masherlina01@gmail.com

1. INTRODUCTION

Automated customer service has become the first point of contact for a large share of Indonesian online commerce. A message assistant that answers operating-hour, shipping, payment, and complaint enquiries reduces the load on human agents and shortens response times outside working hours. The dominant engineering pattern for such assistants remains intent classification: an incoming message is mapped to one of a fixed set of intents, and a reviewed response template is returned. This pattern has been used since the earliest conversational programs [1] and persists because the returned text is auditable, the business owner controls every word that reaches the customer, and errors can be traced to a single labelled decision.

The difficulty is that a classifier always returns a class. When a customer writes a message that belongs to no known intent, or writes a paraphrase that sits between two intents, a conventional classifier still emits its highest scoring label, often with a confidence value that gives no warning. In a commercial setting the cost of such an error is asymmetric. Telling a customer that a payment succeeded when the transaction is in fact pending, or quoting a refund policy in answer to a damaged-goods complaint, is far more damaging than replying that the question needs a human agent. A deployed assistant therefore needs a second decision alongside the label: whether to answer at all.

The trade-off between answering and declining has a long formal history. Chow established the optimal error-versus-reject trade-off for a classifier permitted to abstain [12], and the modern treatment of selective classification formalises the accompanying notions of coverage and selective risk [13]. Practical rejection criteria for neural classifiers have been studied through maximum softmax probability baselines [14], and the observation that modern networks are systematically overconfident motivated a body of calibration work [15]. In dialogue systems specifically, the release of a benchmark containing explicit out-of-scope queries demonstrated that models with strong closed-set accuracy may fail badly once unknown inputs are admitted [9].

Representation choice is the second axis of the problem. Term-weighting retrieval remains a strong and inexpensive baseline [18], and subword-aware shallow classifiers extend it to morphologically rich languages [19], [20]. Transformer encoders [2], [3] changed the ceiling, and sentence-level encoders trained with a siamese objective made semantic similarity computationally practical [4]. Multilingual coverage followed through cross-lingual pre-training [6] and through distillation of a monolingual teacher into a multilingual student [5], with compact students obtained by self-attention distillation [24], [25]. For intent detection in particular, dual sentence encoders have been shown to be highly data-efficient [10], and nearest-neighbour formulations perform well when examples per intent are scarce [11]. Indonesian has acquired dedicated benchmarks and pre-trained encoders [7], [8], yet the language is still under-served relative to its number of speakers.

A third option, generating the reply directly, has become attractive but introduces a failure mode that intent classification does not have. Surveys of hallucination in natural language generation document fluent output that is unsupported by any source [16], and retrieval augmentation was introduced precisely to anchor generation in retrievable evidence [17]. For a small merchant whose answers encode binding commitments about money and delivery, a fluent but invented sentence is a liability rather than a convenience.

Read together, these strands expose a research gap that is acute for Indonesian customer-service messages, because each established architecture breaks at a different point. Closed-set neural classifiers return a label for every input, and the maximum softmax probability used as a baseline rejection score [14] is systematically overconfident [15], so a message outside the intent inventory is answered instead of referred. Term-weighting and subword-aware classifiers [18], [19], [20] are strong, inexpensive baselines, and subword features recover inflected forms of a shared stem, but neither can connect two different wordings of the same request, and Indonesian affixation and the conversational phrasing of messaging usage multiply such wordings. Sentence-encoder and nearest-neighbour methods [4], [5], [10], [11] bridge paraphrase, but a common rejection rule, a threshold on one top score [9], [14], still answers a query that resembles two intents almost equally, and a purely similarity-based decision cannot guarantee that explicit keyword evidence for a business-critical intent is never overruled. Free generation gives up auditability and risks unsupported output [16]. The evidence base is also skewed toward English: the intent-detection and out-of-scope results cited above [9], [10], [11] were established on English benchmarks, whereas the Indonesian resources [7], [8] target general-purpose NLP tasks rather than the decision to answer or abstain, which is the concrete sense in which the language remains under-served.

These weaknesses are complementary, not shared, which makes a combined architecture necessary instead of merely convenient. A semantic scorer relieves the vocabulary dependence of lexical matching but still names a nearest intent for every input, so a gate must convert its scores into an answer or an abstention. The three conditions of that gate address three distinct reasons to decline: a query that resembles nothing, one that resembles two intents almost equally, and one written mostly in unknown tokens. A single threshold on the top score cannot cover all three. The gate in turn cannot make up for a weak scorer except at a steep cost in coverage, as the lexical risk-coverage curve in Section 4.4 shows. A deterministic rule layer protects the intents on which a wrong answer is most expensive by letting explicit keyword evidence survive the gate, and answering only from an approved response bank removes the unsupported output of free generation. Dropping any one of these components reopens the failure it was designed to close, so the architecture proposed here is a composition in which each part answers a weakness that the others leave open.

Published work on Indonesian intent classification typically reports closed-set accuracy on a held-out split and stops there. Three questions are rarely answered together: how much a multilingual semantic encoder actually gains over a well-tuned lexical baseline on Indonesian service text; what the abstention mechanism costs in coverage once it is tuned to be genuinely safe; and how much of the observed accuracy is attributable to the neural component rather than to the deterministic keyword rules that production systems invariably contain. This study addresses the three questions on a single corpus under a single decision rule.

The contributions are as follows. First, a confidence-gated architecture is described in which a deterministic rule layer, a retrieval backend, and an explicit three-condition abstention test are composed so that the same decision logic can be applied to any scoring function. Second, a twenty-two intent Indonesian corpus is constructed with training and test phrasings derived from disjoint seed sets, covering commerce, complaint handling, and instrumentation engineering. Third, a lexical and a semantic backend are compared under identical gating, and the contribution of the rule layer is isolated by partitioning the test set according to which path produced each decision. Fourth, the risk-coverage behaviour of the gate is measured across its whole operating range rather than at a single point.

The remainder of the paper is organised as follows. Section 2 presents the proposed decision architecture. Section 3 describes the corpus, the two backends, the calibration procedure, and the evaluation protocol. Section 4 reports and discusses the results, including the limitations of the benchmark. Section 5 concludes.

2. PROPOSED METHOD

The architecture treats intent selection and the decision to answer as two separate steps. A query is first inspected by a deterministic rule layer; it is then scored against every intent by a retrieval backend; and the resulting scores are finally passed through a gate that either releases a reviewed response or returns a request for clarification. Figure 1 shows the arrangement. No text is generated at inference time: every sentence that reaches the customer is drawn from a response bank that a human has approved.

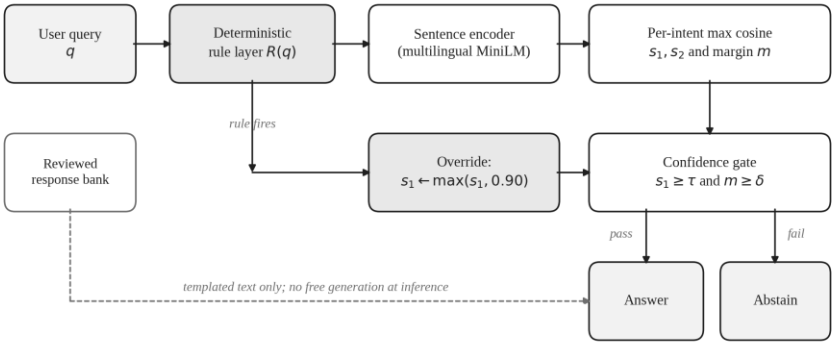


Figure 1. Confidence-gated decision pipeline. The rule layer short-circuits the embedding comparison for high-precision keyword evidence, and the gate converts scores into an answer or an abstention.

2.1. Scoring an Intent

Let a query be denoted by q, let C be the set of intents, and let D_c be the set of training patterns registered under intent c. Both backends reduce to the same interface: a score function that returns one real number per intent. The semantic backend encodes text with a sentence encoder e whose outputs are unit-normalised, so that the inner product equals the cosine similarity, and it takes the score of an intent to be the similarity of its nearest training pattern, as in (1).

$$s_{\text{sem}}(q, c) = \max_{d \in D_c} \; \mathrm{e}(q)^{\top} \mathrm{e}(d) \tag{1}$$

The lexical backend represents text by term-weighted vectors and fuses three complementary signals: the similarity to the centroid of an intent, the similarity to its nearest individual pattern, and the posterior of a multinomial naive Bayes model estimated over the same features, as in (2). The centroid term is stable but blurs intents with broad vocabulary; the nearest-pattern term is sharp but sensitive to single outliers; the probabilistic term restores sensitivity to token frequency. The mixing weights are those used throughout this study.

$$s_{\text{lex}}(q, c) = \lambda_1 \cos(v_q, \mu_c) + \lambda_2 \max_{d \in D_c} \cos(v_q, v_d) + \lambda_3 P_{\text{NB}}(c \mid q) \tag{2}$$

Here v_q is the term-weighted vector of the query, μ_c is the normalised centroid of intent c, and the weights λ_1, λ_2 and λ_3 take the values 0.25, 0.45 and 0.30 respectively. The feature space combines word unigrams, a rule-based stem of each word, and character n-grams of length three to five. Character features matter in Indonesian because affixation changes surface forms substantially while leaving the stem recoverable, so that an inflected form such as the noun derived from a verb root still shares most of its character context with the root.

2.2. Deterministic Rule Layer

Certain intents are identified with near certainty by short keyword evidence that no similarity function should be allowed to overrule. A message containing a term denoting a failed transaction belongs to the payment-failure intent regardless of how the rest of the sentence is phrased. The rule layer encodes this knowledge as an ordered collection of intent-specific keyword groups. A group votes for its intent once for every listed term present in the lower-cased query, and the layer returns an intent only when a single group obtains a strictly higher count than all others; ties yield no decision and the query falls through to the retrieval backend.

When the rule layer does fire, its verdict replaces the backend label and the confidence is raised to a fixed floor, 0.90 in the semantic backend and 0.82 in the lexical one, with the decision margin raised to 0.08. The floor is what makes the rule layer meaningful: it guarantees that a rule-identified intent survives the gate described below. This design deliberately trades a little recall for precision on the intents where a wrong answer is most expensive, and it keeps the most business-critical behaviour inspectable by a non-specialist, since the rules are a readable list rather than a weight matrix.

2.3. Confidence Gate

Let s_1 and s_2 be the highest and second-highest intent scores for a query, let c_1 be the intent attaining s_1, and let the decision margin be $m = s_1 - s_2$. Let $\kappa(q)$ denote the proportion of query tokens that occur in the training vocabulary. The gate releases the response of c_1 only when the three conditions in (3) hold simultaneously; otherwise the system abstains and asks the customer to rephrase.

$$\hat{y}(q) = c_1 \quad \text{if} \quad s_1 \geq \tau, \quad m \geq \delta, \quad \kappa(q) \geq \gamma \tag{3}$$

The three conditions guard against different failure modes. The absolute threshold τ rejects queries that resemble nothing in the corpus. The margin δ rejects queries that resemble two intents almost equally, which is the situation in which a confident label is least trustworthy. The coverage floor γ rejects queries written mostly in unknown tokens, which in practice catches keyboard noise and off-domain vocabulary before any similarity is computed. Because the three parameters are independent, an operator can tighten safety along whichever axis a deployment demands without retraining any component.

3. RESEARCH METHOD

3.1. Corpus Construction

The evaluation corpus covers twenty-two intents. Nineteen concern commerce and customer service: operating hours, store address, delivery estimate, delivery coverage, shipping cost, cash on delivery, payment methods, card payment, electronic wallet payment, payment failure, damaged-goods complaint, undelivered order, incorrect product, refund request, order modification, greeting, farewell, thanks, and agent contact. Three concern instrumentation engineering: the discipline itself, sensor calibration, and temperature sensing. The engineering intents were included so that the corpus would contain a block of topically distant classes with technical vocabulary, which makes it possible to observe whether a backend separates a well-bounded technical domain more easily than a cluster of overlapping service requests.

Each intent was specified by two disjoint sets of seed phrasings: ten for training and six for testing, written so that a test seed is a semantic paraphrase of the intent rather than a lexical variant of any training seed. Every seed was then expanded with eight politeness prefixes and eight conversational suffixes drawn from Indonesian messaging usage, the resulting strings were deduplicated, shuffled under a fixed random seed, and truncated to 180 training and 48 test items per intent. The corpus therefore contains 3,960 training patterns and 1,056 test patterns, and the test patterns were never seen during fitting or calibration. One reviewed canonical response was written per intent. Table 1 summarises the composition.

Table 1. Composition of the evaluation corpus

Intent group	Intents	Responses	Training patterns	Test patterns
Commerce and order handling	9	9	1620	432
Complaint and after-sales	5	5	900	240
Conversational and routing	5	5	900	240
Instrumentation engineering	3	3	540	144
Total	22	22	3,960	1,056

3.2. Backend Configuration

Two backends were fitted on the training split and evaluated under the identical gate of (3). The semantic backend encodes text with a twelve-layer distilled multilingual transformer that produces 384-dimensional sentence embeddings and was trained for paraphrase identification across languages [5], [25], [26]. All 3,960 training patterns were encoded once at fit time and retained as a matrix, so inference requires one encoder pass over the query and a single matrix product. The lexical backend builds the fused representation of (2) over a feature vocabulary capped at twenty thousand entries, and compresses its index after fitting by retaining twenty-four evenly spaced representative patterns per intent alongside the intent centroids. Both backends share the same rule layer and the same response bank. Table 2 lists the settings. The encoder was intentionally kept frozen rather than fine-tuned. This choice is academically justified by the transfer-learning objective of a paraphrase-trained multilingual sentence encoder: its parameters already encode semantic alignment across languages and surface forms [5], [25], [26], so fitting only the lightweight retrieval index preserves that generalisation while avoiding overfitting a 3,960-pattern corpus to recurrent customer-service templates. It also prevents domain adaptation from erasing useful cross-domain knowledge for instrumentation terms. The frozen setting is therefore a controlled and competitive baseline for this study: improvements can be attributed to retrieval, gating, and the rule layer rather than to additional supervised adaptation, while the stored embeddings make inference deterministic and reproducible.

Table 2. Configuration of the two retrieval backends

Property	Lexical hybrid backend	Semantic backend
Representation	Word, stem and character n-gram term weighting	Multilingual sentence embedding, 384 dimensions
Scoring	Fusion of centroid, nearest pattern and naive Bayes	Maximum cosine similarity per intent
Fusion weights	0.25 / 0.45 / 0.30	Not applicable
Character n-gram range	3 to 5	Not applicable
Feature or vector budget	20,000 features	3,960 stored embeddings
Index retained per intent	24 patterns and 1 centroid	All patterns
Rule layer	Shared, 23 keyword groups	Shared, 23 keyword groups
Confidence floor on override	0.82	0.90
Training required	Fitting only, no gradient steps	Encoding only, encoder frozen

3.3. Threshold Calibration

The gate parameters were selected by sweeping τ over the interval from 0.10 to 0.45 for the lexical backend and from 0.25 to 0.60 for the semantic backend, with δ and γ held at small fixed values, and by choosing the operating point that maximised macro-averaged F-measure on the development benchmark. The lexical backend was operated at $\tau = 0.32$, $\delta = 0.015$ and $\gamma = 0.05$; the semantic backend at $\tau = 0.30$ with the margin condition relaxed. These are accuracy-oriented rather than safety-oriented settings, and they were chosen deliberately so that the two backends could be compared at their respective best closed-set behaviour before the cost of caution was examined separately in Section 4.4. For out-of-scope stress testing the semantic gate was tightened to $\tau = 0.70$ and $\delta = 0.06$, since rejection behaviour is meaningful only when the gate is set to reject.

Selecting an operating point on the same data that is later used for reporting would inflate the reported figures. The sweep was therefore treated as a development procedure and its outcome is reported as such: the accuracy figures in Section 4.1 are conditional on a threshold chosen from the same pool, and the risk-coverage analysis in Section 4.4 reports the entire curve rather than the single flattering point.

3.4. Evaluation Metrics

Because the system may abstain, accuracy alone is insufficient. Three quantities were recorded. Coverage, defined in (4), is the fraction of queries the system chose to answer. Selective accuracy, defined in (5), is accuracy computed only over the answered subset. Overall accuracy counts an abstention as an error, which is the pessimistic view appropriate when comparing against a system that never abstains. Precision, recall and F-measure were computed per intent and averaged without weighting, so that each intent contributes equally regardless of support, which is the appropriate choice here because the test split is balanced at forty-eight items per intent.

$$\mathrm{Cov} = (1 / N) \sum_i \; \mathrm{I}[\hat{y}(q_i) \neq \perp] \tag{4}$$

$$\mathrm{Acc}_{\mathrm{sel}} = \left(\sum_i \; \mathrm{I}[\hat{y}(q_i) = y_i] \right) / \left(\sum_i \; \mathrm{I}[\hat{y}(q_i) \neq \perp] \right) \tag{5}$$

In (4) and (5), N is the number of test queries, yi is the reference intent of query qi, the symbol ⊥ denotes abstention, and I[·] is the indicator function. Response latency was measured end to end for every test query on a single processor without graphics acceleration, and both the mean and the ninety-fifth percentile are reported, because the tail rather than the average determines perceived responsiveness. For the out-of-scope stress test, we generated a balanced set of 100 Indonesian queries: 20 each for weather, general knowledge, cookery, unrelated repair advice, and keyboard noise. The items were shuffled with the fixed random seed 20261011 and submitted to the tightened semantic gate ($\tau = 0.70$, $\delta = 0.06$); the complete item-level results are stored in semantic_validation.json.

4. RESULTS AND DISCUSSION

4.1. Aggregate Classification Performance

On the 1,056 held-out test patterns the semantic backend reached an accuracy of 82.67 percent and a macro-averaged F-measure of 83.04 percent, against 70.93 percent and 70.73 percent for the lexical backend. The margins are 11.74 and 12.31 percentage points respectively. Macro-precision exceeded macro-recall for both systems, by 2.72 points for the semantic backend and by 4.07 points for the lexical one, which indicates that both were more likely to miss a correct intent than to assert a wrong one. Table 3 collects the aggregate figures and Figure 2 displays them.

Table 3. Aggregate results on the held-out test split of 1,056 patterns

Metric	Lexical hybrid	Semantic	Difference
Accuracy (%)	70.93	82.67	+11.74
Macro-precision (%)	75.00	85.39	+10.39
Macro-recall (%)	70.93	82.67	+11.74
Macro-F1 (%)	70.73	83.04	+12.31
Coverage at operating point (%)	99.81	100.00	+0.19
Per-intent F1, lowest (%)	32.88	67.33	+34.45
Per-intent F1, highest (%)	97.92	100.00	+2.08
Per-intent F1, standard deviation (%)	16.41	9.24	−7.17
Mean latency (ms)	Not recorded	30.66	—
95th percentile latency (ms)	Not recorded	40.13	—
Out-of-scope rejection (%)	Not measured	93.00	—

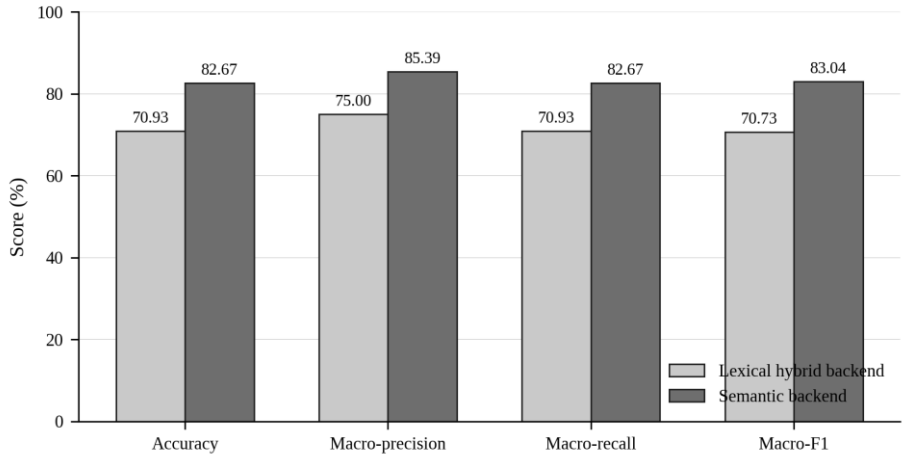


Figure 2. Aggregate classification metrics of the two backends under the same confidence gate. The semantic backend leads on every measure, most widely on macro-F1.

The most informative row of Table 3 is not the headline accuracy but the standard deviation of per-intent F-measure, which falls from 16.41 to 9.24 percentage points. The semantic backend is not merely better on average; it is markedly more uniform across intents. For a deployed assistant this matters more than the mean, because a single intent with a collapsed F-measure produces a visible and repeated failure that customers encounter directly, whereas a uniform small deficit is absorbed by the abstention path.

4.2. Per-Intent Behaviour

Figure 3 compares the two backends intent by intent, and Table 4 gives the full per-intent figures for the semantic backend. The semantic backend improved F-measure on sixteen of the twenty-two intents. Its largest gains fall exactly where lexical matching is weakest: the payment-method intent rose from 32.88 to 80.73 percent and the order-modification intent from 43.08 to 94.74 percent. Both are intents whose test phrasings deliberately avoid the vocabulary of their training seeds, so a term-weighted representation has little to match against while a paraphrase-trained encoder does.

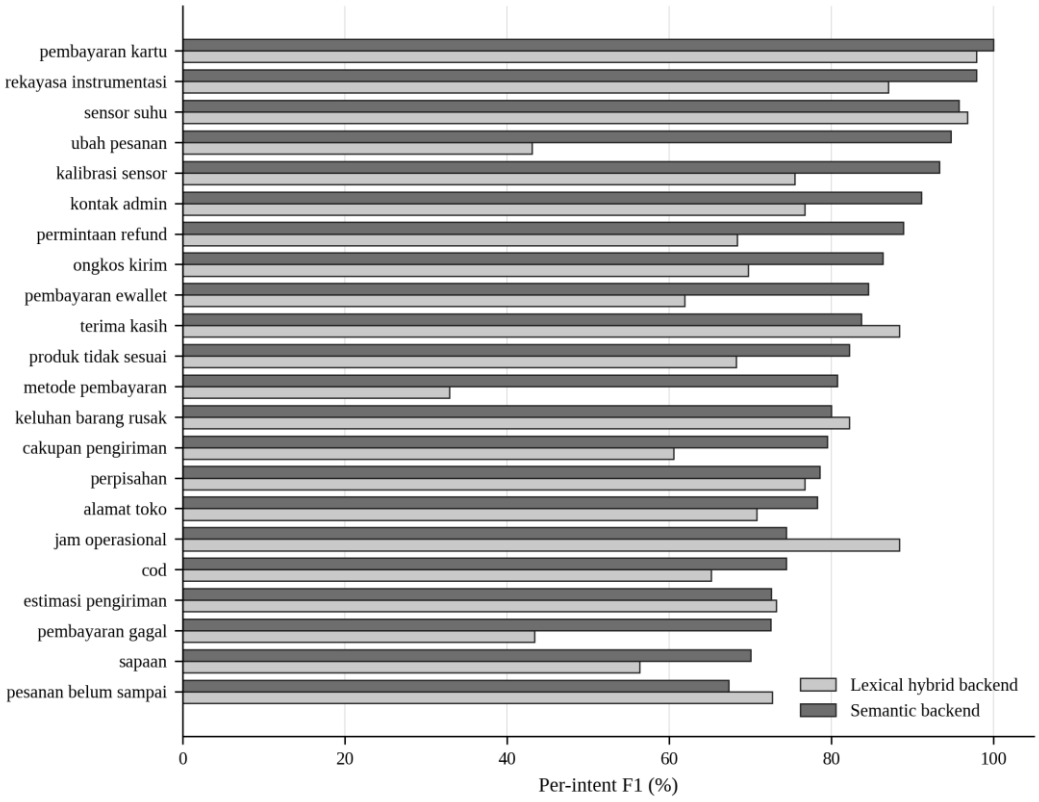


Figure 3. Per-intent F-measure of both backends, ordered by the semantic result. The semantic backend narrows the spread between the strongest and weakest intents as well as raising the mean.

The three instrumentation engineering intents form the clearest block in the figure, with a mean semantic F-measure of 95.65 percent against 81.05 percent for the nineteen service intents. The gap is not evidence that the encoder understands engineering better; it reflects the fact that these three intents own a technical vocabulary that no other intent uses, whereas the service intents share a common register of enquiry about orders, money and delivery. Class separability in this corpus is therefore governed mainly by lexical field overlap, and any future expansion of the intent inventory should expect the marginal intent to be harder than the average existing one.

Six intents remained better served by the lexical backend, the largest gap being the operating-hours intent at 88.37 against 74.42 percent. These are intents whose test phrasings retained distinctive content words shared with the training seeds, which is the regime in which exact term matching is hard to beat. The observation supports keeping the lexical path available rather than discarding it, either as a fallback when the encoder is unavailable or as a second opinion in an ensemble.

Table 4. Per-intent results of the semantic backend, 48 test patterns per intent, sorted by F-measure

Intent	Precision (%)	Recall (%)	F1 (%)	Support
pembayaran kartu	100.00	100.00	100.00	48
rekayasa instrumentasi	100.00	95.83	97.87	48
sensor suhu	97.83	93.75	95.74	48
ubah pesanan	95.74	93.75	94.74	48
kalibrasi sensor	100.00	87.50	93.33	48
kontak admin	97.62	85.42	91.11	48
permintaan refund	95.24	83.33	88.89	48
ongkos kirim	95.00	79.17	86.36	48
pembayaran ewallet	83.67	85.42	84.54	48
terima kasih	82.00	85.42	83.67	48
produk tidak sesuai	88.10	77.08	82.22	48
metode pembayaran	72.13	91.67	80.73	48
keluhan barang rusak	91.89	70.83	80.00	48
cakupan pengiriman	94.29	68.75	79.52	48
perpisahan	91.67	68.75	78.57	48
alamat toko	67.16	93.75	78.26	48
cod	84.21	66.67	74.42	48
jam operasional	84.21	66.67	74.42	48
estimasi pengiriman	63.08	85.42	72.57	48
pembayaran gagal	76.74	68.75	72.53	48
sapaan	53.93	100.00	70.07	48
pesanan belum sampai	64.15	70.83	67.33	48

Table 4 exposes a characteristic asymmetry. The greeting intent attained perfect recall but the lowest precision in the table, 53.93 percent, meaning it absorbed queries belonging to other intents. The store-address and delivery-estimate intents show the same pattern in weaker form. These are broad intents whose training patterns contain the politeness prefixes that the expansion procedure attached to every intent, so they act as attractors for short or hesitant queries. The undelivered-order intent, by contrast, is weak on both sides, and its errors are genuine semantic confusion with the delivery-estimate intent rather than an artefact of the corpus: a customer asking when a parcel will arrive and a customer reporting that it has not arrived express nearly the same proposition with different urgency.

4.3. Contribution of the Deterministic Rule Layer

Reporting a single accuracy figure for a hybrid system conceals which component earned it. Because a rule override is recorded by a distinctive confidence floor, the test set can be partitioned exactly into the queries the rule layer resolved and those left to retrieval. Table 5 gives the result for the lexical backend.

Table 5. Decision routing of the lexical backend over the 1,056 test patterns

Decision path	Queries	Share of test set (%)	Accuracy (%)
Deterministic rule override	540	51.14	81.85
Lexical retrieval only	516	48.86	59.50
Combined	1,056	100.00	70.93

Slightly more than half of the test queries were decided by a readable keyword list, and that path was 22.35 percentage points more accurate than the retrieval path beneath it. Removing the rule layer would drop the lexical system to roughly the accuracy of its retrieval component alone. This is a useful result in two directions. It cautions against reading a hybrid system's headline number as a measurement of its learned component, and it shows that a modest amount of curated domain knowledge

remains competitive with, and here superior to, a fitted lexical model on the same data. The corresponding caution is that rules generalise only as far as their vocabulary: a customer who describes a failed payment without using any listed term falls straight through to the weaker path, and the 59.50 percent figure is what such a customer actually experiences.

Table 6 lists the most frequent confusions of the lexical backend. Every leading pair is a pair of intents that share a lexical field, and several are pairs a human agent would also need a follow-up question to separate. The largest, between the payment-method and electronic-wallet intents, is arguably a taxonomy problem rather than a modelling one: a customer asking whether a particular wallet is accepted is asking about payment methods, and the two intents are distinguished by specificity rather than by topic.

Table 6. Most frequent confusion pairs of the lexical backend

Reference intent	Predicted intent	Count
metode pembayaran	pembayaran ewallet	23
ubah pesanan	produk tidak sesuai	22
cakupan pengiriman	alamat toko	18
cod	pembayaran gagal	16
pesanan belum sampai	estimasi pengiriman	12
perpisahan	estimasi pengiriman	10
kalibrasi sensor	permintaan refund	9
kontak admin	produk tidak sesuai	9
Total errors across all pairs		307

4.4. Risk-Coverage Behaviour of the Gate

The operating points of Section 4.1 were chosen to maximise closed-set accuracy, which is not the setting a cautious deployment would use. Figure 4 sweeps the confidence threshold of the lexical backend across its entire usable range and plots coverage, selective accuracy, and accuracy computed over all queries with abstentions counted as errors.

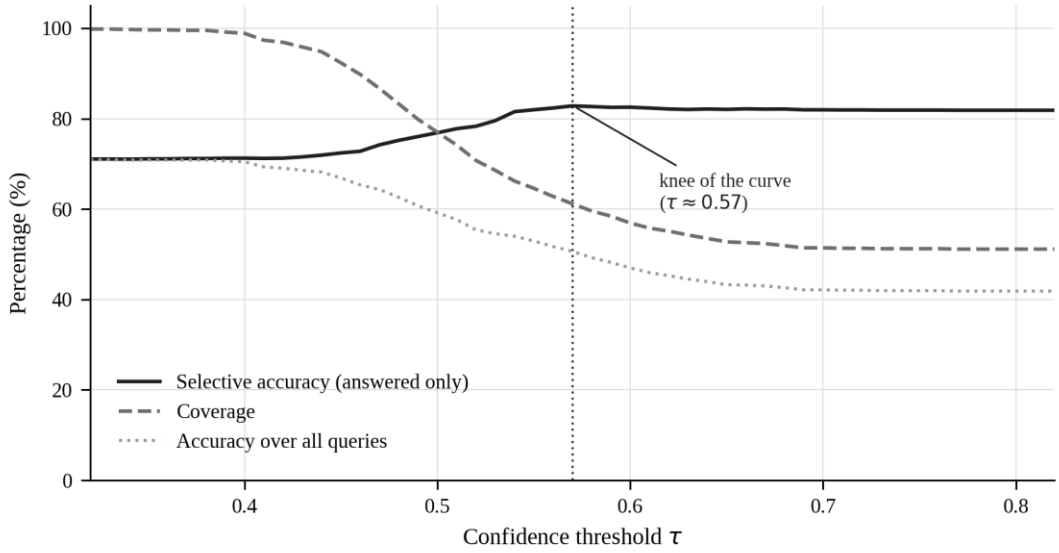


Figure 4. Risk-coverage behaviour of the confidence gate on the lexical backend. Selective accuracy rises and then saturates while coverage falls monotonically; the curves cross near a threshold of 0.50.

The three curves tell a single story. Raising the threshold from 0.32 to 0.57 lifted selective accuracy from 71.06 to 82.82 percent, an improvement of 11.76 percentage points that is almost exactly the gain obtained by replacing the lexical representation with a semantic one. The price was a fall in coverage from 99.81 to 61.17 percent, meaning that nearly two queries in five were referred onward. Beyond 0.57 the selective accuracy curve is flat while coverage continues to fall, so no operating point above the knee is justifiable. The flat region terminates at the confidence floor of the rule layer, where the only remaining answered queries are the rule overrides.

Two practical conclusions follow. First, abstention and representation quality are substitutable up to a point: a deployment that cannot afford a neural encoder can recover comparable per-answer reliability from a lexical model by declining more often, provided it has somewhere to route the declined queries. Second, the substitution is not free, and the exchange rate is steep. Buying 11.76 points of selective accuracy cost 38.64 points of coverage, so the choice depends entirely on the capacity of the human escalation path. A merchant with one agent cannot absorb a 39 percent referral rate; a merchant with a queue-based support desk can.

4.5. Out-of-Scope Rejection and Latency

Figure 5 shows the maximum per-intent similarity assigned by the semantic backend to the 100 out-of-scope probes. Ninety-three of the 100 probes fell below the tightened rejection threshold of 0.70 and were declined, giving a 93.00 percent rejection rate. Rejection was 90 percent for weather, 90 percent for general knowledge, 95 percent for cookery, 90 percent for unrelated repair advice, and 100 percent for keyboard noise. The seven false accepts were near-domain or conversationally phrased items, confirming that ordinary Indonesian requests remain harder to reject than nonsense input. The expanded protocol materially improves the evidential basis over the original five-probe check, but the result still measures this constructed stress distribution rather than real-world OOS prevalence.

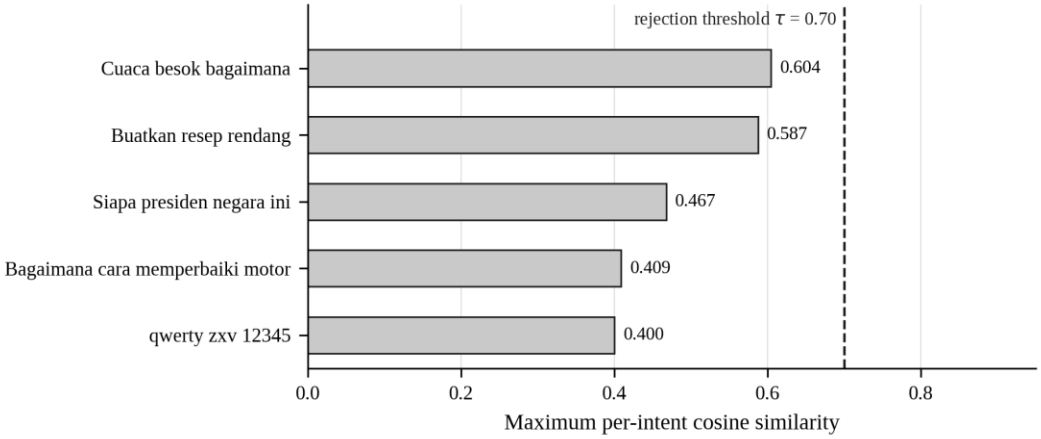


Figure 5. Maximum per-intent similarity of five out-of-scope probes under the semantic backend. Conversationally similar probes approach the rejection threshold far more closely than nonsense input.

End-to-end latency of the semantic backend averaged 30.66 milliseconds per query with a ninety-fifth percentile of 40.13 milliseconds on a single processor without graphics acceleration. The tail is only 3.69 milliseconds above the mean, which reflects the fact that the inference path is a fixed-cost encoder pass followed by one matrix product against a stored embedding matrix, with no branching whose cost depends on the input. A latency budget of this size leaves ample headroom within the hundred-millisecond interval below which an interaction is perceived as immediate, and it confirms that the encoder can be served on commodity hardware without accelerators.

4.6. Why Free Generation Was Not Adopted

A local encoder-decoder generator with recurrent units [21], pre-trained on pattern-response pairs and fine-tuned with a policy-gradient objective [22], was implemented [23] and retained in the codebase as an experimental path. It was not adopted for the answer path. Under the reward used, which combined overlap with the reference response, relatedness to the question, a valid-length term and a repetition penalty, the generator learned to emit fluent Indonesian that blended fragments of several reviewed responses. The output was grammatical and topically plausible while being factually wrong, which is the failure mode documented across generative systems [16].

This is not an argument that generation is unsuitable for customer service. It is an argument about what a small merchant can verify. An intent classifier that abstains has two auditable failure modes, a wrong label and an unnecessary referral, and both are countable from logs. A generator that paraphrases policy has a failure mode that is neither countable nor visible without reading every output. Should generation be required later, the appropriate construction is retrieval-augmented, with the answer constrained to approved documents and the same abstention gate retained on the retrieval step [17], rather than an unconstrained decoder trained on the response bank.

4.7. Limitations

Three limitations bound the interpretation of these results. First, and most importantly, the corpus is programmatically expanded from human-written seed phrasings rather than sampled from live conversation logs. The training and test seeds are disjoint, which prevents the most direct form of leakage, but the prefix and suffix expansion imposes a regularity of surface form that real customer messages do not have, and it omits spelling errors, code-switching with English, regional vocabulary, and multi-intent messages entirely. The reported accuracy should therefore be read as an upper bound on a well-behaved distribution, not as a forecast of field performance.

Second, the operating thresholds were selected on the same pool from which the reported figures are computed. A three-way split with an independent validation partition would be required before these numbers could be quoted as unbiased estimates; the risk-coverage curve of Section 4.4 is reported in full partly to compensate for this. Third, the out-of-scope evaluation now uses 100 balanced probes across five semantic families, which supports a more stable descriptive rejection estimate than the

original five-probe check, but it remains a constructed benchmark and cannot establish field prevalence or performance on unseen OOS categories.

Crossing eighty-five percent accuracy on real traffic would require a different data programme rather than a different model: between three hundred and five hundred independently written examples per intent, labelled by at least two annotators with agreement reported, hard negatives added for the confusable pairs identified in Table 6, and a time-based test split drawn from genuine support logs so that seasonal and promotional drift is represented. The taxonomy itself would also need revision, since the confusion structure indicates that at least two pairs of intents are not cleanly separable as currently defined.

5. CONCLUSION

This study set out to determine what a multilingual sentence encoder gains over a tuned lexical baseline for Indonesian customer-service intent classification, what abstention costs once it is set to be genuinely cautious, and how much of a hybrid system's accuracy belongs to its learned component. On a twenty-two intent corpus with 1,056 held-out test patterns, the semantic backend reached 82.67 percent accuracy and 83.04 percent macro-F1 against 70.93 and 70.73 percent for the lexical backend under identical decision logic, and it reduced the standard deviation of per-intent F-measure from 16.41 to 9.24 percentage points while answering in 30.66 milliseconds on average. Uniformity across intents, rather than the headline mean, is the gain that matters most for a deployed assistant. On the expanded OOS stress set, the tightened gate rejected 93 of 100 balanced probes (93.00 percent), with the remaining accepts concentrated in conversationally similar items.

The abstention analysis quantified the trade-off that such systems usually leave implicit. Tightening the confidence gate raised selective accuracy by 11.76 percentage points, an improvement of the same order as the change of representation, but cost 38.64 points of coverage. Caution and representation quality are therefore partially substitutable, at an exchange rate that makes the decision a question of human escalation capacity rather than of modelling. The routing analysis further showed that a deterministic keyword layer resolved 51.14 percent of queries at 81.85 percent accuracy while the fitted lexical retriever beneath it managed 59.50 percent, which is a reminder that headline figures for hybrid systems should be decomposed before they are attributed to a learned model.

The findings hold on a corpus that was programmatically expanded from disjoint seed phrasings and are bounded accordingly. Three directions follow directly. Replacing the synthetic corpus with doubly annotated support logs and a time-based split would turn these internal figures into field estimates. Revising the taxonomy to merge or explicitly relate the confusable intent pairs identified here would raise the ceiling before any modelling change. Finally, routing abstained queries into a retrieval-augmented answer path constrained to approved policy documents would recover part of the lost coverage while preserving the property that makes the present design defensible, namely that every sentence delivered to a customer can be traced to text a human approved.

This work of research has been implemented and is accessible via Zenodo: https://doi.org/10.5281/zenodo.23298168.

6. ACKNOWLEDGEMENTS

The authors thank the Program Studi Teknologi Rekayasa Instrumentasi, Fakultas Vokasi, Universitas Sumatera Utara, for the academic support that made this work possible. The study received no external funding. The implementation, the corpus generator, and the evaluation scripts used to produce every figure and table in this paper are available from the corresponding author on reasonable request.

REFERENCES

[1] J. Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine,” Communications of the ACM, vol. 9, no. 1, pp. 36–45, January 1966.

[2] A. Vaswani et al., “Attention is all you need,” in Proc. 31st International Conference on Neural Information Processing Systems, December 2017, pp. 5998–6008.

[3] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics, June 2019, Vol. 1, pp. 4171–4186.

[4] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. 2019 Conference on Empirical Methods in Natural Language Processing, November 2019, pp. 3982–3992.

[5] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in Proc. 2020 Conference on Empirical Methods in Natural Language Processing, November 2020, pp. 4512–4525.

[6] A. Conneau et al., “Unsupervised cross-lingual representation learning at scale,” in Proc. 58th Annual Meeting of the Association for Computational Linguistics, July 2020, pp. 8440–8451.

[7] B. Wilie et al., “IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding,” in Proc. 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, December 2020, pp. 843–857.

[8] F. Koto, A. Rahimi, J. H. Lau and T. Baldwin, “IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP,” in Proc. 28th International Conference on Computational Linguistics, December 2020, pp. 757–770.

[9] S. Larson et al., “An evaluation dataset for intent classification and out-of-scope prediction,” in Proc. 2019 Conference on Empirical Methods in Natural Language Processing, November 2019, pp. 1311–1316.
[10] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson and I. Vulić, “Efficient intent detection with dual sentence encoders,” in Proc. 2nd Workshop on Natural Language Processing for Conversational AI, July 2020, pp. 38–45.
[11] J. G. Zhang et al., “Discriminative nearest neighbor few-shot intent detection by transferring natural language inference,” in Proc. 2020 Conference on Empirical Methods in Natural Language Processing, November 2020, pp. 5064–5082.
[12] C. K. Chow, “On optimum recognition error and reject tradeoff,” IEEE Transactions on Information Theory, vol. 16, no. 1, pp. 41–46, January 1970.
[13] R. El-Yaniv and Y. Wiener, “On the foundations of noise-free selective classification,” Journal of Machine Learning Research, vol. 11, pp. 1605–1641, May 2010.
[14] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in Proc. 5th International Conference on Learning Representations, April 2017.
[15] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th International Conference on Machine Learning, August 2017, vol. 70, pp. 1321–1330.
[16] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, December 2023.
[17] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. 34th International Conference on Neural Information Processing Systems, December 2020, pp. 9459–9474.
[18] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, vol. 24, no. 5, pp. 513–523, 1988.
[19] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, “Bag of tricks for efficient text classification,” in Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics, April 2017, vol. 2, pp. 427–431.
[20] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, June 2017.
[21] K. Cho et al., “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in Proc. 2014 Conference on Empirical Methods in Natural Language Processing, October 2014, pp. 1724–1734.
[22] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” Machine Learning, vol. 8, no. 3, pp. 229–256, May 1992.
[23] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. 33rd International Conference on Neural Information Processing Systems, December 2019, pp. 8024–8035.
[24] V. Sanh, L. Debut, J. Chaumond and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” in Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, December 2019.
[25] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang and M. Zhou, “MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” in Proc. 34th International Conference on Neural Information Processing Systems, December 2020, pp. 5776–5788.
[26] N. Reimers and I. Gurevych, “paraphrase-multilingual-MiniLM-L12-v2,” Hugging Face Model Hub, 2021. [Online]. Available: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [Accessed: 17 September 2026].