

MOSAIC: Automated Intent Description Generation for Multi-Label Classification

Ajay Chatterjee Eric Crestan Raunak Saha
Rotem Mordoch Swarnangsu Acharyya
Microsoft

{ajchatterjee, ericres, raunaksaha,
Rotem.Mordoch, swarna}@microsoft.com

Abstract

Traditional intent classifiers map utterances directly to intent categories using labeled examples. In contrast, LLM-assisted zero-shot intent classifiers rely on natural language descriptions alongside intent labels to determine the most appropriate categories for a given utterance. As intent taxonomies grow and evolve, descriptions must define clear boundaries among overlapping intent categories. Maintaining them is costly because it requires domain expertise and continual updates to keep pace with changing product and user behavior. We present MOSAIC (Multi-label Ontology Synthesis for Automated Intent Classification), a deployed system that automatically generates and updates intent descriptions for existing taxonomy. MOSAIC analyzes unlabeled user utterances to identify coherent clusters. It associates cluster level evidence with the existing intent taxonomy. Then, It synthesizes discriminative descriptions that define each category’s semantic boundary. On a 140-class production deployment, MOSAIC approaches the quality of expert written descriptions while reducing taxonomy revision time from 2-4 weeks to less than 30 minutes. On the public NLU++ and Multi³NLU++ benchmark datasets, MOSAIC generated descriptions do match or exceed the performance of expert written descriptions.

1 Introduction

In deployed conversational assistants, user utterances provide the most direct signal of evolving user needs. Routing these utterances into a structured intent taxonomy is therefore essential for product analytics, failure triage and feature prioritization. In our deployment, this taxonomy spans 140+ active classes, receives queries in dozens of languages, and changes as the product evolves. Three coupled problems dominate at this scale.

Taxonomy drift. Without an explicit discovery loop, the taxonomy ages out of relevance within a quarter.

Multi-label queries. Realistic user utterances routinely express several co-applicable intents. As a synthetic example:¹

“How do I reset my password and also update the email address on my account?”

This maps to at least four labels: *password_reset*, *account_settings*, *email_update*, and *how*. Clustering would map this to a single cluster, but the combinatorial space of intent co-occurrences makes cluster-based resolution impractical.

Cross-lingual scale. Discovery, description generation and classification must all be language-agnostic by construction.

Recent methods such as ClusterLLM (Zhang et al., 2023) and Dial-In LLM (Hong et al., 2025) use LLM feedback to improve cluster quality and discover intent groups. Other approaches fine-tune smaller models on labelled data (Tunstall et al., 2022; Zhang et al., 2020) or use retrieval-augmented classifiers (Lewis et al., 2020). However, in privacy-constrained deployments where customer data cannot be used for fine-tuning or RAG indexing, these approaches are not viable. In such settings, manually curated intent descriptions are often the only mechanism for steering a zero-shot classifier, yet writing them requires domain expertise that scales with taxonomy size. When the product evolves monthly, manual curation becomes a bottleneck.

We describe MOSAIC, a deployed system that closes this loop for a large-scale production assistant. A discovery component embeds queries with a multilingual encoder, partitions the pool with K-Means clustering, refines partitions with density clustering, and uses two LLM steps under light human review to keep the taxonomy current (§3.1). The same pipeline generates intent descriptions, short natural-language specifications of each category’s scope, from cluster evidence and intent

¹Synthetic; no customer text is reproduced.

labels without relying on human-authored ontology text. An LLM-based multi-label classifier then receives these descriptions alongside intent categories, enabling it to disambiguate semantically overlapping categories in a single LLM call.

We make three contributions. (i) We show that intent descriptions are essential for large-taxonomy classification and that they can be generated automatically from cluster evidence. In production, an initial taxonomy can be bootstrapped from domain knowledge; the system then automatically updates existing descriptions and discovers new intents with descriptions. On the public NLU++ benchmark, generated descriptions significantly outperform expert descriptions on micro-F1 and sample-F1 ($p < 0.05$; macro-F1 trends positive; §4.2). (ii) We ablate the source of improvement. An LLM can rewrite intent labels into descriptions without any cluster evidence, approaching expert quality on public benchmarks. Adding cluster grounding on top yields a further +2.6 pp on NLU++ EN and +1.4 pp on Multi³NLU++. On these public benchmarks, whose short label strings already carry most of the intent semantics, cluster grounding contributes a modest but consistent gain. In our production taxonomy of 140 semantically overlapping intents, where label strings alone are insufficient, cluster grounding contributes substantially more of the total improvement (§4.2). (iii) We release the generated NLU++ and Multi³NLU++ descriptions to motivate further research on this area.

2 Related Work

Intent discovery and taxonomy induction. New-intent discovery has moved from embedding-space methods (Chatterjee and Sengupta, 2020; Zhang et al., 2022; Tang et al., 2024; Zhang et al., 2024) to LLM-in-the-loop approaches. CLUSTER-LLM (Zhang et al., 2023) uses LLM triplet feedback to refine embeddings for clustering; Dial-In LLM (Hong et al., 2025) uses cluster-level binary coherence evaluation to iteratively improve cluster quality and trains a downstream single-label classifier on the result. Shah et al. (2025) use LLMs to generate and validate intent taxonomies from search logs, but focus on taxonomy structure rather than discriminative descriptions for classification. Huang and He (2025) cast text clustering itself as LLM-based classification. These systems address a different problem from ours: they focus on taxon-

omy discovery or cluster quality, whereas MOSAIC assumes an existing taxonomy and uses clustering as a means to **generate intent descriptions** for downstream multi-label classification. This decouples discovery (unsupervised, LLM-free) from classification (description-driven, multi-label).

Intent descriptions for classification. Zero-shot classification via natural-language label descriptions was pioneered by Yin et al. (2019), who frame it as textual entailment. Recent work confirms that richer descriptions substantially improve zero-shot performance: Gao et al. (2023) show that fine-tuning on label descriptions alone (no labelled examples) yields 17–19 pp gains over standard zero-shot models, while Kumar et al. (2024) achieve further improvements with contextualized label descriptions. Parikh et al. (2023) show that short manually written descriptions improve single-label zero-shot accuracy on MASSIVE (60 intents). However, all of these approaches assume **descriptions are manually provided**. As taxonomies grow and evolve, manual description authoring becomes a bottleneck. MOSAIC addresses this gap by automatically generating descriptions from cluster evidence, reducing the need for human authoring. Large language models have shown strong zero-shot classification ability (Brown et al., 2020; Ye et al., 2023), and chain-of-thought prompting (Wei et al., 2022) further improves reasoning on complex tasks. MOSAIC combines these ideas: it uses CoT-prompted LLM classification with automatically generated descriptions. A direct comparison on MASSIVE would require extending it to multi-label annotation; we therefore evaluate on NLU++ and its multilingual extension Multi³NLU++, the largest public multi-label intent benchmarks available. To our knowledge, MOSAIC is the first system to **automatically generate** intent descriptions from data evidence and show that they exceed expert-authored descriptions on a public multi-label benchmark.

Multi-label intent classification. Banking-77 (Casanueva et al., 2020), MASSIVE (FitzGerald et al., 2023), and NLU++/Multi³NLU++ (Casanueva et al., 2022; Moghe et al., 2023) are standard benchmarks; the first two are single-label. NLU++/Multi³NLU++ is, to our knowledge, the largest public multi-label intent benchmark (60+ classes), making it the natural testbed for description-driven classification.

3 System

Terminology. An *intent category* is a semantic class in the taxonomy, while an *intent label* is its name, such as `cancel_close_leave_freeze`. An *intent description* is a short natural-language specification of the category’s scope.

MOSAIC has three components sharing a single compliant ingestion path (Fig. 1): a discovery loop that keeps the taxonomy current (§3.1), a description generator that produces discriminative intent specifications from cluster evidence (§3.2), and a classifier that routes queries against the described taxonomy (§3.3).

3.1 Discovery loop

Multilingual embedding and partitioned clustering. Each query is embedded with a multilingual Sentence-BERT encoder (Reimers and Gurevych, 2019)², so semantically equivalent queries in different languages collapse to nearby points. The embedding pool is first partitioned with K-Means (MacQueen, 1967) for scalability. Within each partition, we then run iterative dbscan, a DB-SCAN extension that uses a relaxed density schedule (Ester et al., 1996). It iteratively extracts dense clusters under successively looser thresholds. This process extracts well-formed clusters first and recovers rare query patterns in later passes. Keeping the LLM out of this stage is critical at production scale. Density-based clustering requires only $O(C)$ downstream LLM calls (where C is the number of discovered clusters) compared to $O(N)$ calls for LLM-in-the-loop methods.

LLM cluster labelling. For each candidate cluster, we randomly sample representative utterances and send them to an LLM for abstraction. The LLM generates a short intent-like label that summarizes the shared user need expressed by the cluster and assigns a categorical confidence score based on cluster coherence. The rubric is applied by the LLM itself under an explicit instruction: **High** indicates a clear dominant intent ($\geq 70\%$ of sampled utterances agree), **Medium** indicates moderate consistency (51–69% agreement), and clusters without a majority intent are assigned `no_intent` and discarded. On NLU++ EN, 343/368 (93.2%) clusters receive High labels and 25/368 (6.8%) receive Medium; on Multi³NLU++ the distribution

is 93.7%/5.8%/0.6%. Random sampling is chosen for scalability under fixed prompt-length budgets; alternative strategies such as medoid-based or diversity-aware selection are natural extensions and we leave a systematic comparison to future work. The confidence label is used *only* as an internal selection gate for the description-generation stage: Low-confidence clusters are filtered out to prevent noisy evidence from influencing the generated descriptions. It is **not** consumed by the classifier, does not appear in intent descriptions, and does not propagate to downstream classification decisions. We use a three-level categorical score rather than a scalar because it acts as a coarse filter, not a calibrated probability; calibrated scalar confidence is an interesting extension we do not pursue in this work.

3.2 Description generation

The cluster evidence accumulated during discovery contains rich information about each intent’s scope. The pipeline starts with **zero human-authored descriptions**: each intent is seeded with its intent category (underscores replaced by spaces, e.g., `cancel_close_leave_freeze` → “cancel close leave freeze”). No expert ontology text enters the loop. We then generate **discriminative intent descriptions** (short natural-language specifications grounded in observed usage) in **two separate LLM calls**. **(1) Cluster-grounded identification:** The first LLM call processes cluster summaries in batches. For each batch, it identifies the intents supported by the cluster evidence and proposes or refines a description for each affected intent. Because a single intent may be supported by many clusters, its description is updated incrementally as evidence accumulates across the full run. **(2) Mutual-exclusivity refinement:** a second LLM call jointly rewrites the minimum subset of descriptions needed to reduce semantic overlap and enforce mutual exclusivity between closely related intents, preventing confusion between semantically adjacent categories. The resulting descriptions are generated entirely automatically. As we show in §4.2, they recover the vast majority of expert quality on production data (micro-F1) and significantly outperform expert descriptions on most metrics on a public benchmark.

Implementation details. The description generator uses GPT-4.1 (long-context deployment) for the main results in Table 1, and Llama-3.1-70B-Instruct for the open-weight evaluation in

²Production encoder: `sentence-transformers/paraphrase-multilingual-mpnet-base-v2`; see Appendix B.

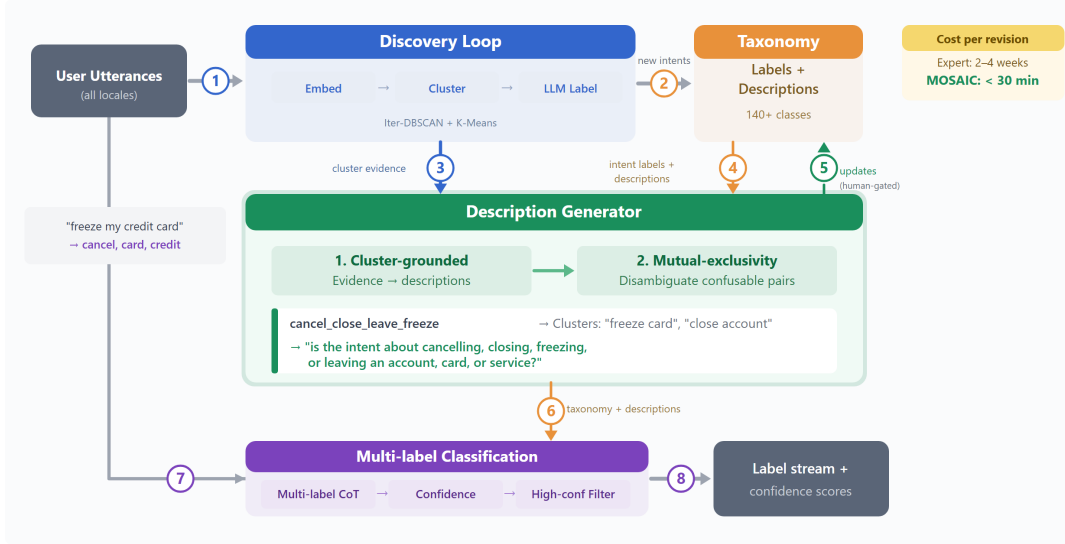


Figure 1: End-to-end MOSAIC flow. The **discovery loop** embeds multilingual queries, clusters with density based clustering algorithm, and uses LLM labelling to identify intents. The **description generator** produces intent descriptions grounded in cluster evidence. The **multi-label classifier** routes queries against the taxonomy using intent labels and descriptions. Taxonomy updates are human-gated. The numbered edges indicate the execution order: ① ingest queries → ② register newly discovered intents in the taxonomy → ③,④ feed cluster evidence and intent labels to the description generator → ⑤ commit refined descriptions back to the taxonomy (human-gated) → ⑥ publish taxonomy + descriptions to the classifier → ⑦ route user queries → ⑧ emit labels with confidence scores.

§4.7. Both use temperature 0.1 and max_tokens = 8,192. The generator sees only **train-fold cluster summaries and intent labels**; test utterances are never seen during clustering or description generation. Note that while the classifier is zero-shot with respect to labelled examples, the description generator uses unsupervised cluster summaries derived from unlabelled in-domain data. The overall pipeline is more accurately characterised as unsupervised description induction rather than purely zero-shot. Each cluster summary contains an abstract label (a short phrase assigned by the LLM) and the cluster’s quality rating; raw utterances are not passed to the description generator. The generator’s direct contribution from clusters, is mapping intent labels to the correct super-category (e.g., deciding whether **international** should be categorised under “events or actions” vs. “items, services, or accounts”). The two-pass pipeline requires approximately $N_{\text{batches}} + 1$ LLM calls. Generated descriptions are concise, averaging 9 words in length. Template-based descriptions uniformly follow the form is the intent about <PLACEHOLDER>?, whereas freeform descriptions allow more flexible wording while remaining length constrained.

3.3 Multi-label classification

For each query, a single LLM call receives the full taxonomy (intent categories *and/ or* descriptions) and returns the set of applicable labels with per-label confidence (high / medium / low). Descriptions disambiguate labels that share surface-level names and anchor relevance judgements in cluster-grounded evidence. $|\mathcal{C}| \approx 140$ labels and description prompt fits within the model’s context window, eliminating the need for a retrieval stage.

4 Experiments

4.1 Production evaluation

We evaluate on a held-out, human-labelled judgement set of $N = 1,652$ production queries stratified across $K = 140$ active taxonomy classes including the long tail (54 classes with ≤ 5 gold items, 46 with 6–20, 22 with 21–50, 15 head classes with ≥ 51 , and 3 with no evaluated items). The set is genuinely multi-label: gold annotations average 2.01 labels per query. Ten trained expert reviewers with product context performed the annotation; annotators saw 10 LLM-generated candidate labels per query and could additionally add any missing label from the full taxonomy. We measured consistency on a 100-query subset and observed

an approximately 2:1 agreement-to-disagreement ratio. Full annotation protocol is described in Appendix A.

4.2 Description quality: the central result

We evaluate description quality on two settings: the public NLU++ (Casanueva et al., 2022) and Multi³NLU++ (Moghe et al., 2023) benchmarks (62 intents, multi-label, four languages), and a held-out production judgement set (140 classes, $\sim 1,650$ queries). In all settings, we compare description sources: intent category only, expert written description, and MOSAIC generated description; holding the classifier prompt and LLM fixed. We compare two description writing styles. The template based style is motivated by the question style descriptions used in NLU++ and Multi³NLU++ dataset. The freeform style is length constrained and reflects the description format used in our production taxonomy. Table 1 reports the results for all description sources and evaluation settings. We use micro-averaged precision, recall, and F1 as primary metrics because they weight every (query, label) prediction equally, reflecting overall system accuracy in production routing where each label assignment triggers a downstream action. We complement these with macro-averaged precision and recall to verify that gains are not concentrated on frequent intents.

Public benchmarks. On NLU++ English, generated descriptions with the question-template outperform expert descriptions on micro-F1 (77.9 vs. 75.2), and paired bootstrap tests confirm significance (Appendix C). The key difference is that expert descriptions are written generically from domain knowledge (e.g., “is the intent to ask about something related to a transfer, payment or deposit?”), while cluster-grounded descriptions are anchored in observed usage patterns. This grounding makes them more discriminative, substantially improving precision (76.8% vs. 69.8%) with only a small recall trade-off (79.2% vs. 81.7%). On the four-language Multi³NLU++ pool, the same English-induced descriptions transfer across languages with consistent gains over intent label only baselines.

We evaluate on the non-empty gold-label subset of NLU++ EN ($N = 264$ of 302), because standard multi-label metrics are undefined on queries with no gold labels. Consequently our numbers are not directly comparable to prior work that reports on

Dataset	Configuration	Micro			Macro		
		P	R	F1	P	R	$\Delta F1$
NLU++ ($N = 264$)	<i>GPT-4.1 backbone</i>						
	intent label only	73.1	73.7	73.4	77.5	85.0	—
	Label + LLM desc. (from label)	71.5	79.5	75.3	78.8	85.3	+1.9
	Label + LLM desc. (clusters, template)	76.8	79.2	77.9	80.6	87.6	+4.5
	Label + LLM desc. (clusters, freeform)	76.9	68.8	72.6	82.1	81.1	+0.8
	Label + expert desc. (human)	69.8	81.7	75.2	79.1	86.2	+1.8
	<i>Llama-3.1-70B-Instruct backbone</i>						
	intent label only	37.2	62.3	46.6	43.6	72.3	—
	Label + LLM desc. (from label)	53.9	66.6	59.6	64.8	77.3	+13.0
	Label + LLM desc. (clusters, template)	53.6	70.1	60.7	64.6	79.3	+14.1
Multi ³ NLU++ (4 lang., $N = 1,056$)	Label + LLM desc. (clusters, freeform)	54.2	65.5	59.3	61.2	76.2	+12.7
	Label + expert desc. (human)	54.5	67.9	60.5	64.9	78.5	+13.9
	<i>GPT-4.1 backbone</i>						
	intent label only	71.4	71.8	71.6	73.6	81.7	—
	Label + LLM desc. (from label)	70.3	78.6	74.2	76.2	82.8	+2.6
	Label + LLM desc. (clusters, template)	72.3	79.1	75.6	76.4	84.3	+4.0
	Label + LLM desc. (clusters, freeform)	75.1	64.8	69.6	80.0	74.3	+2.0
	Label + expert desc. (human)	69.8	80.0	74.6	77.1	83.7	+3.0
	<i>Llama-3.1-70B-Instruct backbone</i>						
	intent label only	31.2	51.9	39.0	37.9	59.6	—
Internal ($K = 140$, 10 lang., $N = 1,652$)	Label + LLM desc. (from label)	43.4	54.5	48.3	56.3	61.1	+9.3
	Label + LLM desc. (clusters, template)	48.3	50.7	52.3	58.8	63.2	+13.3
	Label + expert desc. (human)	47.7	57.7	52.2	60.0	64.2	+13.2
	<i>GPT-4.1 backbone</i>						
	intent label only	43.1	41.0	42.0	38.6	41.5	—
	Label + LLM desc. (from label)	46.7	39.9	43.0	43.3	38.9	+1.0
	Label + LLM desc. (clusters, template)	50.6	42.5	46.2	42.4	38.9	+4.2
	Label + LLM desc. (clusters, freeform)	50.6	49.1	49.8	41.2	41.6	+7.8
	Label + expert desc. (human)	53.7	48.1	50.7	44.7	42.1	+8.7

Table 1: Description-source ablation across public and internal benchmarks. All rows within a backbone use the same classifier prompt and LLM; only the intent descriptions differ. Public benchmark results are computed on items with non-empty gold labels ($N = 264$ of 302 for NLU++ EN). $\Delta F1$ = micro-F1 gain over intent label only baseline. Bold indicates best per column within each (dataset, backbone) block. Significance tests for the Llama backbone are reported in Table 2.

the full release, and expert descriptions are taken *verbatim* from the NLU++ dataset release without modification. Our goal is to isolate the effect of description source under a fixed zero-shot classifier rather than to establish a new state of the art; fine-tuned systems with access to labelled training data remain above our zero-shot ceiling.

Production. The benefit of descriptions grows with task complexity. On NLU++ (62 intents, short utterances), intent labels alone already achieve reasonable performance. On production data (140 classes, complex multi-intent queries, $\sim 1,650$ evaluated), the gap between intent label only ($\mu F1 = 42.0$) and expert descriptions ($\mu F1 = 50.7$) widens substantially, confirming that descriptions are essential for complex tasks.

4.3 Template vs. freeform descriptions

Description format interacts with taxonomy scale. We compare **template** (is the intent about <PLACEHOLDER>?) vs. **freeform** (unconstrained natural language) styles. At 62 intents (NLU++), the template wins on F1: $\mu F1$ 77.9 vs. freeform 72.6. However, freeform achieves the **highest micro-precision** across configurations

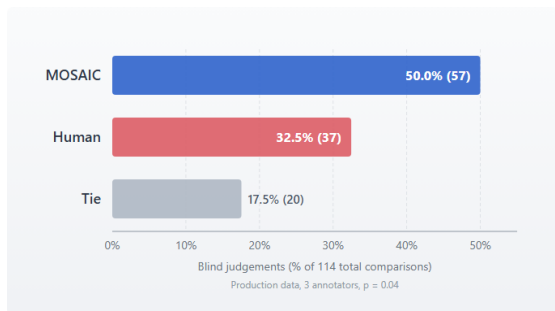


Figure 2: Blind side-by-side adjudication over 114 judgments with three reviewers. Reviewers selected the better label set without knowing which system produced it.

(76.9%, matching template’s 76.8%), with a recall trade-off (68.8% vs. 79.2%). On Multi³NLU++, the same pattern holds: freeform precision (75.1%) exceeds template (72.3%) but recall drops (64.8% vs. 79.1%). At 140 classes (production), the pattern reverses: freeform recovers most of the expert gap (49.8 vs. 50.7), while template falls behind (46.2). The rigid template aids discrimination at small scale with simpler class definition but becomes restrictive at large scale, where the LLM needs expressive freedom. Table 7 provides the full ablation across five configurations: intent labels alone, LLM-generated template descriptions from intent labels, LLM-generated template descriptions from cluster evidence, LLM-generated freeform descriptions from cluster evidence, and expert-written descriptions. This design separates the effect of description format from the effect of cluster grounding.

Per-intent analysis. Generated descriptions deliver the largest gains on ambiguous multi-system names (e.g., `international` +26.7 pp, `cancel_close_leave_freeze` +5.2 pp) and regress on 9 of 62 simple names where the super-category phrase introduces confusion with neighbours.

4.4 Side-by-side adjudication

Gold-label noise. Automatic metrics against a fixed gold set can conflate system errors with missing or incomplete gold labels (Northcutt et al., 2021; Plank, 2022). Among the 145 labels counted as false positives for the MOSAIC-generated descriptions, 71% were also predicted by the same classifier when using expert-written descriptions. This suggests that many scored false positives are plausible labels missing from the gold set rather

than clear system errors.

Blind comparison. In a blind side-by-side study, three annotators compared the label sets produced using generated descriptions and expert written descriptions. Annotators preferred the labels produced with generated descriptions in 51.9% of judgments on NLU++ EN ($p = 0.73$) and 60.6% on production data ($p = 0.04$; Fig. 2). A consensus-corrected evaluation, in which a predicted label is added to the gold set when two of the three systems agree on it, raises micro-F1 from 77.9 to 87.5 (Table 4). Because all systems share the same LLM backbone, this figure is best interpreted as an upper bound on the effect of incomplete annotation rather than as a corrected estimate of true performance; the +2.4 pp gap between generated and expert descriptions is stable across the original and corrected gold sets.

4.5 Discovery validation

We validate clustering on Banking-77, CLINC-150, and MASSIVE (Appendix H). With SBERT, MOSAIC reaches 81–92% of the best published NMI from LLM-in-the-loop methods while keeping clustering entirely LLM-free. Importantly, SBERT recovers most gold intents as majority labels of at least one cluster (72/77 on Bank-77, 150/150 on CLINC, 54/60 on MASSIVE), confirming that the cluster evidence for nearly every intent is presented to the LLM labeller. In production, the monthly cycle discovers 5–15 candidates, of which ~ 2 survive human review.

4.6 Deployment and operational context

MOSAIC has processed 200K+ classified queries over 12 months. Classification uses GPT-4.1 (long-context deployment) in a single LLM call ($\sim 3,500$ input tokens, 2–5s p95 latency). The end-to-end cost per taxonomy revision includes: compute for embedding and clustering the query pool, followed by ~ 40 LLM calls for description generation ($\sim \$0.50$ in LLM API cost). The total pipeline completes in under 30 minutes, replacing 2–4 weeks of manual curation.

In production, MOSAIC serves as a **human-gated taxonomy management** tool. The pipeline proposes new candidate intents with generated descriptions and updates descriptions for existing intents, while domain experts review and approve changes before they enter the live taxonomy. This workflow reduces expert effort from full ontology

Comparison (NLU++, Llama)	Δ	95% CI	p
Label-derived – Label-only ($\mu F1$)	+13.0	[+10.2, +15.8]	<.001***
MOSAIC – Label-derived (Recall)	+3.5	[+1.0, +6.1]	.007**
MOSAIC – Label-derived (Precision)	−0.3	[−3.0, +2.3]	.79
MOSAIC – Expert ($\mu F1$)	+0.3	[−2.5, +3.1]	.86

Table 2: Paired bootstrap significance tests (10K resamples) on the NLU++ open-weight reproduction. MOSAIC refers to the cluster-grounded template variant, which is the strongest Llama-authored configuration in Table 1.

curation to targeted review, enabling the taxonomy to keep pace with a rapidly evolving product and user needs.

4.7 Open-weight validation

To confirm the results are not tied to a single proprietary LLM, we reproduce the full pipeline (description generation and classification) using Llama-3.1-70B-Instruct (Grattafiori et al., 2024), an open-weight, self-hostable model. Numerical results appear in the Llama backbone blocks of Table 1; the cluster-grounded template variant of MOSAIC is the best-performing description source on both NLU++ (60.7) and Multi³NLU++ (52.3), slightly exceeding expert-written descriptions in both cases. The freeform variant was only reproduced on NLU++ EN because multilingual Llama freeform descriptions were not generated in this study; on NLU++ it trails the template variant (59.3 vs. 60.7 $\mu F1$), consistent with the GPT-4.1 finding that freeform underperforms on smaller public benchmarks. Paired bootstrap significance tests (Table 2) show that (i) descriptions provide the dominant improvement (+13.0 pp, $p < .001$); (ii) cluster grounding contributes a significant +3.5 pp recall gain over label-derived descriptions ($p = .007$) with no precision loss; and (iii) MOSAIC (template) is statistically indistinguishable from expert-written descriptions ($p = .86$).

5 Conclusion

MOSAIC is a deployed system for human-gated taxonomy management in large-scale intent classification. It automates manual ontology curation by generating and updating intent descriptions from cluster-level evidence. Rather than replacing expert judgment, MOSAIC shifts expert effort from manual ontology authoring to reviewing and updating generated candidate intents and descriptions. We evaluate description quality through downstream classification against expert-written descriptions

and multiple generated-description baselines, including descriptions generated from intent labels alone.

Our experiments show that high-quality descriptions are important for reliable large-taxonomy classification. Automatically generated descriptions consistently outperform intent-label-only classification. They also approach or exceed the quality of expert-written descriptions across both public benchmarks and production settings. We find that cluster grounding provides useful signal beyond what an LLM can infer from intent labels alone. We also find that the best description format depends on taxonomy scale. Template-based descriptions work best on smaller public benchmarks, while freeform descriptions are more effective in the larger production taxonomy.

In production, MOSAIC reduces taxonomy revision time from an estimated 2-4 weeks to under 30 minutes. The full pipeline: embedding, clustering, LLM-based cluster labeling, description generation and refinement runs entirely inside a compliant environment. This makes MOSAIC a practical solution for maintaining evolving intent taxonomies. Experts remain in control of taxonomy changes, while the system automates the costly process of generating and updating discriminative intent descriptions.

Limitations

Our evaluation includes one production deployment with a large intent taxonomy, so the results may not fully generalize to other products, domains or smaller taxonomies. Although we evaluate on public benchmarks, these datasets are smaller than our production setting and may not capture the same taxonomy drift, multilingual traffic and multi-intent behavior observed in deployment.

The quality of the expert-description baseline is another limitation. Some NLU++ intent labels and expert descriptions are relatively generic, so a more carefully optimized human-written baseline might narrow the gap between generated and expert-written descriptions. In addition, our strongest production comparison combines cluster grounding with freeform description style, so it does not fully separate the contribution of cluster evidence from the contribution of description format. We include a controlled 2×2 factorial ablation on public benchmarks to isolate these factors, but an equivalent full factorial ablation on production data remains future

work.

Description style also interacts with taxonomy scale. Template-based descriptions perform best on the public benchmarks with 62 intents, but they underperform freeform descriptions in the production taxonomy with 140 intent classes. This suggests that a single fixed description format may not be optimal across taxonomies and motivates adaptive formatting as future work.

The primary results in Table 1 rely on a proprietary LLM (GPT-4.1), which limits full third-party reproducibility. We mitigate this by reproducing the pipeline on an open-weight Llama-3.1-70B-Instruct backbone (§4.7), where the same qualitative ordering holds and MOSAIC descriptions remain statistically indistinguishable from expert-written descriptions. Extending the reproduction to additional open-weight families (e.g., Qwen) remains valuable future work. Finally, we evaluated MOSAIC only on intent classification. The workflow itself is task-agnostic: any downstream task guided by natural-language label descriptions (entity typing, slot filling, topic classification, product categorization, hierarchical classification) could plausibly benefit from usage-grounded description generation. We leave empirical validation on such tasks to future work. We have not evaluated robustness with weaker base models, and the classifier confidence scores remain uncalibrated.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. [Efficient intent detection with dual sentence encoders](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.
- Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. [NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics.
- Ajay Chatterjee and Shubhashis Sengupta. 2020. [Intent mining from past conversations for conversational agent](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4140–4152, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, page 226–231. AAAI Press.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Nataraajan. 2023. [MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.
- Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. [The benefits of label-description training for zero-shot text classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13823–13844, Singapore. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Mengze Hong, Wailing Ng, Chen Jason Zhang, Yuanfeng Song, and Di Jiang. 2025. [Dial-in LLM: Human-aligned LLM-in-the-loop intent clustering for customer service dialogues](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5885–5900, Suzhou, China. Association for Computational Linguistics.
- Chen Huang and Guoxiu He. 2025. [Text clustering as classification with llms](#). In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2025*, page 374–384, New York, NY, USA. Association for Computing Machinery.
- Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2024. [Gen-z: Generative zero-shot text classification with contextualized label descriptions](#). In *The Twelfth International Conference on Learning Representations*.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.
- Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2023. [Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics.
- Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. [Pervasive label errors in test sets destabilize machine learning benchmarks](#). *Preprint*, arXiv:2103.14749.
- Soham Parikh, Mitul Tiwari, Prashil Tumbade, and Quaizar Vohra. 2023. [Exploring zero and few-shot techniques for intent classification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 744–751, Toronto, Canada. Association for Computational Linguistics.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, Tara Safavi, Siddharth Suri, Mengting Wan, Leijie Wang, and Longqi Yang. 2025. [Using large language models to generate, validate, and apply user intent taxonomies](#). *ACM Trans. Web*, 19(3).
- Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. [Learning geometry-aware representations for new intent discovery](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5641–5654, Bangkok, Thailand. Association for Computational Linguistics.
- Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. [Efficient few-shot learning without prompts](#). *Preprint*, arXiv:2209.11055.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Junjie Ye, Xuanning Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhua Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. [A comprehensive capability analysis of gpt-3 and gpt-3.5 series models](#). *Preprint*, arXiv:2303.10420.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.
- Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. [Discriminative nearest neighbor few-shot intent detection by transferring natural language inference](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.
- Shun Zhang, Chaoran Yan, Jian Yang, Wei Zhang, Changyu Ren, Tongliang Li, Jiaqi Bai, and Zhoujun Li. 2024. Tinid: A transfer and interpretable llm-enhanced framework for new intent discovery. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 195–212, Cham. Springer Nature Switzerland.
- Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. [ClusterLLM: Large language models as a guide for text clustering](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics.
- Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022. [New intent discovery with pre-training and contrastive learning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 256–269, Dublin, Ireland. Association for Computational Linguistics.

A Annotation protocol

For the internal production evaluation, the judgment set was annotated by 10 trained reviewers with product context. For each query, annotators were shown 10 LLM-generated candidate intent labels and asked to mark all applicable candidates. They then inspected the full taxonomy to add any missing labels not present in the candidate set. We measured annotation consistency on a 100-query subset, obtaining an approximately 2 : 1 agreement-to-disagreement ratio.

For the side-by-side evaluation study, reviewers were shown the same query together with two unordered sets of predicted intent labels. One is from the generated-description system and another one is from the expert-description system. The source of each label set was hidden. Reviewers selected the better label set or marked the comparison as a tie when both outputs were equally appropriate.

B Clustering configuration

The clustering procedure runs in five outer stages with distance thresholds 0.12, 0.25, 0.30, 0.40, 0.50 and corresponding min-samples values 25, 10, 10, 8, 8. Within each stage, the inner loop runs for at most 100 iterations. We set $\delta_{\text{dist}} = 0.01$, $\delta_{\text{min-samples}} = -1$, and an oversize-cluster cap of 500 utterances.

C Bootstrap CIs and significance tests

All intervals use 10 K bootstrap resamples; p -values are paired bootstrap (same resampled indices for both systems).

Metric	LLM (clusters, templ.)	Expert (human)	p (LLM > Exp.)
<i>NLU++ EN</i> ($N = 264$)			
Exact match	.367 [.311,.428]	.307 [.254,.364]	.017*
Micro-F1	.779 [.751,.806]	.752 [.726,.778]	.006**
Macro-F1	.817 [.746,.829]	.799 [.725,.808]	.062
Sample-F1	.764 [.733,.795]	.737 [.706,.768]	.015*
<i>Multi³NLU++</i> ($N = 1,056$)			
Exact match	.330 [.301,.360]	.288 [.261,.315]	<.001***
Micro-F1	.756 [.741,.770]	.746 [.732,.759]	.032*
Macro-F1	.781 [.761,.794]	.777 [.758,.790]	.280
Sample-F1	.743 [.727,.759]	.732 [.716,.747]	.025*

Table 3: Bootstrap 95% CIs and paired p -values. MO-SAIC-generated descriptions significantly outperform expert on micro-F1, sample-F1 ($*p < .05$, $**p < .01$, $***p < .001$).

D Consensus-corrected evaluation

Automatic evaluation against a fixed gold set can penalize plausible labels that are missing from the annotation. To estimate the effect of this noise, we construct a consensus-corrected gold set for NLU++ EN. A predicted label is added to the gold set when at least two of the three systems agree on it. This correction raises micro-F1 for all systems, but the gap between generated and expert descriptions remains stable at +2.4pp. Since all systems share the same LLM backbone, this analysis should be interpreted as an upper bound rather than a replacement for human re-annotation.

Configuration	Original μ F1	Corrected μ F1	Δ
LLM (clusters, template)	77.9	87.5	+9.6
Expert (human)	75.2	85.1	+9.8
intent label only	73.4	83.1	+9.7

Table 4: NLU++ EN micro-F1 under the original and consensus-corrected gold sets.

E Per-language classification

Table 5 breaks down Multi³NLU++ micro-F1 by language. English-generated descriptions transfer consistently across all four languages, with cluster-grounded template descriptions achieving the best micro-F1 in three of four languages and within 0.2 pp of expert on Amharic. Amharic shows the smallest gains, consistent with its typological distance from English.

Config	English	Spanish	Marathi	Amharic	All
intent label only	73.9	69.9	73.7	69.0	71.6
LLM desc. (from label)	75.3	73.9	76.2	71.7	74.2
LLM desc. (clusters, template)	77.3	76.1	77.3	71.7	75.6
LLM desc. (clusters, freeform)	71.3	69.2	70.6	67.3	69.6
Expert desc. (human)	75.3	74.3	76.8	71.9	74.6

Table 5: Per-language micro-F1 on Multi³NLU++. Cluster-grounded template descriptions match or outperform expert in three of four languages (Amharic is a tie within 0.2 pp). Freeform descriptions consistently underperform intent-label-only across all languages.

F Description examples

Table 6 shows representative intent descriptions generated from cluster evidence. We compare expert descriptions with template-based and freeform generations to illustrate how the two styles differ in specificity and naturalness.

# Intent	Cluster Evidence	Expert	Generated (template)	Generated (freeform)
1 cancel close leave freeze	C1: cancel appointment C2: cancel reservation C3: cancel loan C4: cancel room booking ... (10 clusters)	is the intent asking about canceling or closing something?	is the intent about cancel, close, leave, or freeze items, services, or accounts?	Requests to cancel, close, leave, freeze, or block an account, card, item, booking, or service.
2 international	C1: ask about atm withdrawal abroad C2: ask about using bank services abroad	is the intent to ask about something related to international issues?	is the intent about international items, services, or accounts?	Confirming or stating that something is international, such as a card, payment, or transfer.
3 transfer payment deposit	C1: inquire payment status C2: inquire payment timing C3: make payment to person C4: make top up (5 clusters)	is the intent to ask about something related to a transfer, payment or deposit?	is the intent about transfer, payment, or deposit events or actions?	Requests or actions to transfer, send, or deposit money between accounts or to a person.
4 how much	(no clusters)	is the intent asking about some quantity or how much?	is the intent about how much amounts, prices, or values?	Questions about cost, price, or amount for services, items, or events.
<i>Generated description regresses</i>				
5 more higher after	(no clusters)	is the intent to indicate something more, higher, after or increasing?	is the intent about more, higher, or after amounts, prices, or values?	

Table 6: Description examples on NLU++ with cluster evidence. Template descriptions preserve all underscore stems and anchor to super-categories; freeform descriptions are more natural but variable. Intents with cluster evidence (top 3) get more specific descriptions; those without (bottom 2) rely on name-only inference. The last row shows a regression where the template narrows scope to financial amounts.

G Description ablation

We isolate the contributions of the question-template format and cluster grounding with a 2×2 factorial ablation (Table 7).

Configuration	NLU++ EN		Multi ³ NLU++	
	μ F1	P / R	μ F1	P / R
intent label only	73.4	73.1 / 73.7	71.6	71.4 / 71.8
+ LLM desc. (clusters, freeform)	72.6	76.9 / 68.8	69.6	75.1 / 64.8
+ LLM desc. (label, template)	75.3	71.5 / 79.5	74.2	70.3 / 78.6
+ LLM desc. (clusters, template)	77.9	76.8 / 79.2	75.6	72.3 / 79.1
+ Expert desc. (human)	75.2	69.8 / 81.7	74.6	69.8 / 80.0

Table 7: Description ablation on NLU++ EN and Multi³NLU++ (non-empty gold). Freeform achieves the highest micro-precision (76.9% EN) but lower recall (68.8%), resulting in lower F1 (-0.8 pp EN, -2.0 pp Multi³NLU++ vs. intent-label-only). In precision-critical production settings, freeform may be preferred despite lower F1.

Two findings emerge. First, description format strongly affects the precision-recall trade-off. On the public benchmarks, template-based descriptions achieve the best F1, while freeform descriptions are more precise but substantially less recall-oriented. This suggests that a fixed question template provides useful activation cues for consistent multi-label prediction at smaller taxonomy scale. However, the production results in Table 1 show the opposite pattern: at 140 classes, freeform descriptions outperform the template format. This

indicates that the optimal description format depends on taxonomy scale.

Second, cluster grounding provides signal beyond what an LLM can infer from intent labels alone. Within the fixed-template setting, adding cluster-derived evidence improves micro-F1 from 75.3 to 77.9 on NLU++ EN and from 74.2 to 75.6 on Multi³NLU++. This comparison holds the LLM, prompt structure and description format fixed; the only difference is whether cluster evidence is provided. The consistent gains across both datasets support the claim that cluster grounding contributes useful information for description generation.

H clustering Result

We evaluate intent discovery on public benchmarks using KMeans partitioning with $K = \lceil N/5000 \rceil$ on the train splits, followed by density-based clustering stage. We report NMI, purity and majority-label coverage. Coverage measures whether each gold intent is surfaced by at least one discovered cluster. Formally, coverage is the fraction of gold intents that appear as the majority label of at least one discovered cluster. This metric is useful in our setting because clusters are used as evidence for downstream LLM labeling and description generation, not as the final classifier.

As shown in Table 8, SBERT recovers most gold

intents as majority labels of at least one cluster: 72/77 on Bank-77, 150/150 on CLINC, and 54/60 on MASSIVE. Although MOSAIC does not match LLM-in-the-loop methods on NMI, it keeps clustering LLM-free while still surfacing evidence for nearly every intent.

Dataset	Method	NMI	Pur.	Cov.
Bank-77	ClusterLLM [†]	85.2	–	–
	Dial-In [†]	82.3	–	–
	MOSAIC (mBERT)	47.5	36.8	77/77
	MOSAIC (SBERT)	73.1	75.5	72/77
CLINC	ClusterLLM [†]	94.0	–	–
	Dial-In [†]	94.1	–	–
	MOSAIC (mBERT)	59.7	52.9	150/150
	MOSAIC (SBERT)	76.6	84.1	150/150
MASS.	Dial-In [†]	78.1	–	–
	MOSAIC (mBERT)	52.7	51.1	53/60
	MOSAIC (SBERT)	71.6	80.0	54/60

Table 8: Intent discovery results on public benchmarks. [†] denotes published numbers from prior LLM-in-the-loop clustering methods. Cov. = majority-label coverage.