

From Synthetic to Localized: Benchmarking Multilingual Intent Classification in Logistics Customer Service

Haoyu He^{1,2*} Jinyu Zhuang¹ Haoran Chu^{1,2*}

Shuhang Yu¹ Hao Wang² Kunpeng Han^{1†}

AI Department, J&T Express¹

School of Information Science and Technology, ShanghaiTech University²

{robert.zhuang, wayne.yu, lion.han}@jtexpress.com,

{hehy12024, chuhr2023, wanghao1}@shanghaitech.edu.cn

Abstract

Multilingual intent classification is central to logistics customer-service systems, where models must handle noisy user queries across languages and hierarchical labels. Yet many multilingual benchmarks rely on machine-translated text, which is cleaner than real customer requests and can overestimate robustness. We present a public benchmark for hierarchical multilingual intent classification from real logistics logs: approximately 30K de-identified, single-utterance queries curated from 600K records through filtering, LLM-assisted quality control, human privacy review, and sampled label auditing. The dataset uses a two-level taxonomy with 13 parent and 17 leaf intents; English, Spanish, and Arabic are seen languages, while Indonesian, Chinese, and Thai support zero-shot evaluation. To measure the gap between machine-translated and professionally localized evaluation, we provide paired test sets derived from the same source items. We benchmark multilingual encoders, embedding models, and small language models under flat and hierarchical protocols, and find that machine translation tests yield higher accuracy in most settings.¹

1 Introduction

Intent classification is a core problem in natural language understanding and underpins many interactive systems, including customer-service automation, virtual assistants, and tool-using language models (Larson et al., 2019; Casanueva et al., 2020; Schick et al., 2023). In multilingual customer-service settings, the task is especially challenging:

semantically equivalent intents may be expressed differently across languages and dialects, while real user queries are often short, noisy, domain-specific, and unevenly distributed across labels. On global logistics platforms, systems must route requests about tracking, delivery failures, address changes, refunds, claims, fees, and branch availability, often under hierarchical label taxonomies (Schuster et al., 2019; Gung et al., 2023; Plaud et al., 2024).

Existing multilingual benchmarks only partly reflect these conditions. MTOP and MASSIVE provide broad language coverage, but much of their multilingual data comes from machine translation or controlled collection rather than naturally occurring customer requests (Li et al., 2021; FitzGerald et al., 2023). Customer-service datasets are closer in domain, but often target multi-turn dialogue, sentiment, search, or matching rather than stand-alone hierarchical intent routing (Chen et al., 2021; Gung et al., 2023; Wang et al., 2025). Thus, current benchmarks may underrepresent the lexical variability, shorthand, code-mixing, and long-tail distributions of production traffic (Artetxe et al., 2023; Dutta Chowdhury et al., 2022).

This mismatch also affects evaluation. Machine-translated test sets are typically more standardized than professionally localized text and may therefore yield overly optimistic estimates of multilingual robustness (Artetxe et al., 2023; Dutta Chowdhury et al., 2022). This issue is particularly important for logistics customer service, where small differences in phrasing can change whether a request should be routed to tracking, delivery change, service scope, fee, or claim workflows.

We introduce a public benchmark for multilingual intent classification built from real logistics customer-service logs. The dataset contains approximately 30K de-identified labeled training utterances curated from roughly 600K historical records through rule-based filtering, LLM-assisted quality

^{*}This work was mainly done while the marked authors were interns at J&T Express.

[†] Corresponding author

¹Prompts and sample data are available at <https://github.com/JTExpress-MICCS/MICCS>. A preprint version of this work is available at <https://arxiv.org/abs/2603.23172>.

control, human privacy screening, and a sampled label audit. It is organized into a partially hierarchical taxonomy with 13 parent intents and 17 leaf intents: nine parents map to one leaf, while four parents contain eight leaves. English, Spanish, and Arabic are included as seen languages, while Indonesian, Chinese, and Thai are reserved as test-only languages for zero-shot evaluation. Because the benchmark is derived from naturally occurring, stand-alone user queries used for routing, it preserves the noise and distributional properties of real customer-service traffic.

We also benchmark the gap between machine translation and professional localization, alongside zero-shot transfer, across multilingual encoders, embedding models, and small language models under flat and parent-constrained protocols. Machine translation scores 1.3 accuracy points higher on average across 45 model-language-task settings, with a 95% bootstrap interval of $[0.159, 2.479]$ over test items.

In summary, our contributions are as follows:

- We release a public real-world benchmark for multilingual intent classification in logistics customer service, together with a curation pipeline that combines de-identification, filtering, LLM-assisted quality control, privacy screening, and human quality auditing.
- We benchmark multilingual encoders, embedding-based classifiers, and small language models under parent, flat-leaf, and parent-constrained settings.
- We show empirically that machine-translated evaluation often overestimates model performance relative to professional localization.

2 Dataset Construction

We construct a multilingual intent benchmark from historical customer-service logs. The released data consist of de-identified user queries paired with a practical two-level intent taxonomy.

2.1 Construction Pipeline

Source data. The raw pool comprises roughly 600K customer-service conversations collected in November 2025 from J&T Express operations serving Mexico and Arabic-speaking markets. Removing agent messages and merging consecutive customer turns leaves approximately 120K customer-side utterance units. English, Spanish, and Arabic

are used as seen languages, while Indonesian, Chinese, and Thai are reserved for zero-shot transfer evaluation. The remaining funnel is summarized in Appendix Table 3.

De-identification and cleaning. Before annotation and release, we remove personally identifiable information, such as tracking numbers and detailed addresses, using rule-based filters, a locally deployed LLM privacy judge, and manual item-by-item review of every retained utterance. We then exclude bot-generated templates, exact, near, and keyword-level duplicates, malformed or truncated queries, and utterances without a clear intent. Privacy filtering and deduplication leave roughly 80K candidates; subsequent quality and scope filters yield approximately 30K training examples.

Taxonomy construction. We group the cleaned queries by operational function and user goal, and iteratively refine the label set to balance coverage, label separability, and per-class support. The final taxonomy is partially hierarchical: nine parent categories map one-to-one to a leaf, while four parents divide into eight leaves. Details of the clustering configuration and complete mapping are provided in Appendix B and Table 4.

Annotation and quality control. Training and test annotation follow different procedures. The approximately 30K training examples receive LLM-assisted scope screening and preliminary intent labels. Human annotators independently audit a random 5% training sample (approximately 1,500 items) to detect systematic errors; raw agreement is about 95% and Cohen’s κ is about 0.94. This sampled audit concerns label quality only; privacy screening covered the full retained set. The separately constructed seen-language test set contains approximately 1,300 items, all independently human-labeled, with disagreements adjudicated by a third expert. Construction and semantic-preservation checks for the test-only-language sets are described in Section 3.2.

3 Dataset Characteristics

3.1 Overview

Appendix Table 3 summarizes the corpus statistics. The release contains approximately 30K training queries, together with development data, the independently annotated seen-language test set, and the test-only-language evaluation sets. A representative de-identified example is shown below:

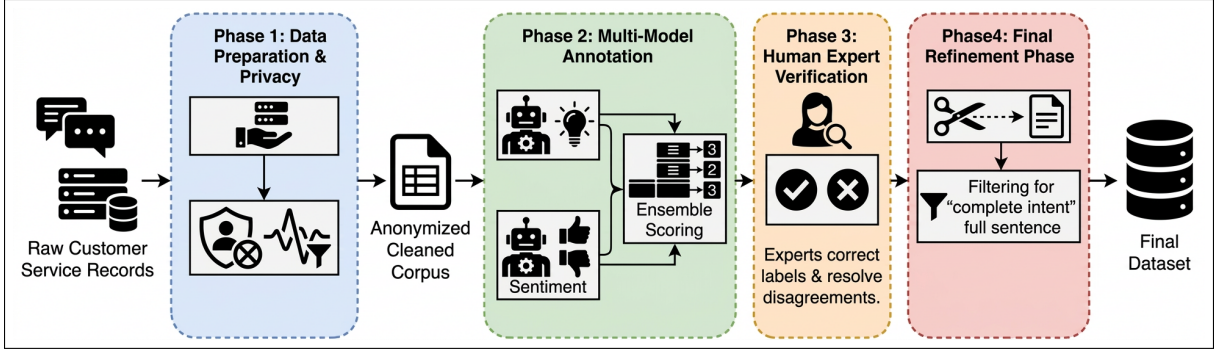
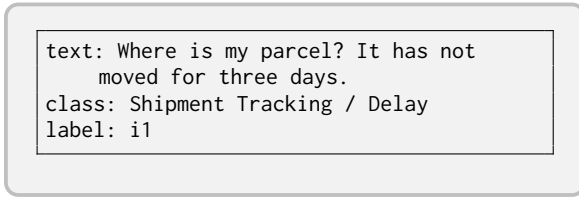


Figure 1: Construction pipeline for the benchmark.



To improve label generation stability in generative classification, we prepend a fixed prefix (e.g., “i1”) to all label IDs.

3.2 Machine-Translated and Human-Localized Evaluation

A key feature of the benchmark is a paired comparison between machine translation (MT) and professional human localization. For each test-only language, Table 2 uses 674 source items. The GPT/Gemini filter is applied once to the shared Spanish source set and keeps items that both systems classify correctly and consistently. Each retained item is then rendered independently by google/translategemma-27b-it (Finkelstein et al., 2026) and by a local business translator, who is asked to preserve the intent while using natural local phrasing. Each pair therefore fixes the source text, label, filtering decision, source language, translation direction, and target language, differing only in its target-language rendering. Spanish was chosen because translation quality was stable in our setting; other source languages were not tested. Below, we call these the machine-translated and human-localized conditions.

The seen-language evaluation in Table 1 instead uses approximately 1.3K naturally occurring English, Spanish, and Arabic test examples. All 17 leaves are represented in the test-only-language sets (7–93 examples per leaf); the three rarest are Shipment Details (7), Fee Dispute / Incorrect Charge (10), and Business Partnership (16). Results for these leaves have greater sampling un-

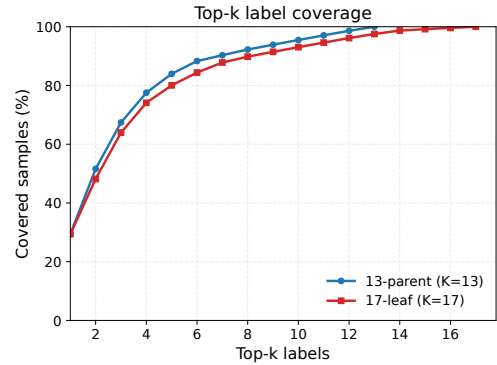


Figure 2: Top- k label coverage for the 13-parent and 17-leaf tasks. Splitting four parent categories produces the additional leaf-level tail.

certainty. Cross-split semantic duplicates are removed.

3.3 Hierarchical Labels, Long-Tail Sparsity, and Noise

The benchmark supports a 13-way parent task and a primary 17-way leaf task (Appendix Table 4). The extra granularity is confined to eight leaves under four parents; the other nine parents are already leaves.

As shown in Figure 2, the 17-way leaf task is more long-tailed than the parent task, consistent with production traffic in which a small number of intents dominate and many low-frequency cases remain sparse. Naturally occurring source queries retain shorthand, misspellings, colloquial phrasing, and cross-lingual variation.

4 Experiments

4.1 Benchmark Tasks and Evaluation Protocol

We evaluate three customer-service routing settings. The 13-way parent task tests coarse operational routing (e.g., tracking vs. fee, claim, or invoice

issues). The 17-way flat leaf task is the main fine-grained setting, requiring direct prediction of operationally similar intents without using hierarchy at inference time. The 17-way parent-constrained task asks generative models to produce a parent intent and a taxonomy-consistent leaf intent. The final evaluation is performed on the leaf label, while the parent prediction constrains the output space.

Unless otherwise stated, all trainable models are trained on the released training split and selected on the development split using only English, Spanish, and Arabic. Each trainable model receives nine hyperparameter configurations, selected on the same development split. Closed-source LLMs are evaluated as few-shot references; several instruction, definition, demonstration, and serialization variants are compared on development data, and the selected prompt is evaluated at temperature 0. All results are reported under the benchmark’s natural traffic distribution, without class rebalancing or synthetic augmentation.

For unseen test-only languages such as Indonesian, Chinese, and Thai, we evaluate zero-shot transfer without target-language supervision, translated training data, or target-language prompt tuning. Evaluation uses the test-only-language conditions defined in Section 3.2.

Our primary metric is accuracy, which is equivalent to Micro-F1 in this single-label classification setting. We use accuracy because the benchmark preserves the naturally imbalanced traffic distribution of real customer-service systems, and overall routing accuracy is the most deployment-relevant objective in this setting. We report Macro-F1 alongside the main seen-language results and provide complete metric tables in Appendix C. All hyperparameters, prompts, and checkpoints are selected on the development split only. Test sets are used strictly for final evaluation.

4.2 Baselines

We compare multilingual encoders, embedding-based classifiers, instruction-tuned small language models, and API models. Most neural models have at most 1B parameters; API models are included only as few-shot references.

For multilingual encoder baselines, we evaluate **mBERT** (Devlin et al., 2019), **XLM-R** (Conneau et al., 2020), and the public **mmBERT** checkpoint jhu-clsp/mmBERT-base (Marone et al., 2026). All use task-specific output layers and the same data access and trial budget; mmBERT uses its

standard tanh-pooled representation without added dropout.

For embedding-based classification, we evaluate **EmbeddingGemma** (Schechter Vera et al., 2025) paired with a lightweight classifier, as well as a **Gemma 3 270M (embedding+classifier)** variant (Gemma Team et al., 2025). Both reuse a frozen encoder, so only the classifier is trained.

We further evaluate instruction-tuned small language models, including **Gemma 3 270M**, **Gemma 3 1B** (Gemma Team et al., 2025), and **Qwen 3 0.6B** (Yang et al., 2025). Unlike the encoder baselines, these accept a new label in the prompt without a change to the output layer.

Finally, we compare against three few-shot prompted closed-source LLMs accessed via API: **Seed-1.8** (Bytedance Seed, 2026), **DeepSeek V3.2** (DeepSeek-AI, 2025), and **Qwen3-Max-Preview** (Alibaba Cloud, 2025).

All systems use the same label inventory, test splits, and evaluation scripts; trainable models additionally use the released train/dev splits for supervised optimization and model selection. The training settings can be found in Appendix E.

5 Results

5.1 Seen-Language Performance

Table 1 reports results on the approximately 1.3K-example naturally occurring English, Spanish, and Arabic test set; following Section 4.1, we report Accuracy/Micro-F1 and Macro-F1. All trainable open models in the table are supervised using the released training split, whereas the closed-source API LLMs use validation-selected few-shot prompts and are included as reference systems rather than fine-tuned baselines.

Across model families, the 13-way parent task is usually easier than the 17-way leaf tasks. The gap comes from the eight leaves that share a parent, where the competing labels are operationally adjacent and unevenly supported.

Detailed Macro-F1 results are provided in Table 5. Difficulty-floor baselines score 5.88–78.54% on the translated 17-way set, while the two lexical baselines fall to 6.23–7.27% under Thai zero-shot transfer; Appendix D gives the full results.

Table 1: Seen-language benchmark results on English, Spanish, and Arabic (Accuracy/Micro-F1, %). Trainable open models are supervised on the released training split; closed-source API LLMs are evaluated with few-shot prompting only.

Model	13-way	17-flat	17-hier
<i>Multilingual encoders</i>			
mBERT	92.62	91.38	N/A
XLNet	94.00	91.85	N/A
mmBERT	94.69	93.15	N/A
<i>Embedding-based classifiers</i>			
EmbeddingGemma	94.85	93.38	N/A
Gemma 3 270M (embedding+classifier)	93.62	92.15	N/A
<i>Instruction-tuned SLMs</i>			
Gemma 3 270M	94.08	93.08	93.23
Qwen 3 0.6B	94.23	93.46	93.77
Gemma 3 1B	95.00	93.92	94.08
<i>Closed-source API LLMs (few-shot prompting; no fine-tuning)</i>			
Seed-1.8	91.69	92.23	91.23
DeepSeek V3.2	93.92	90.92	89.85
Qwen3-Max-Preview	94.77	91.15	92.92

5.2 Real-World Robustness: Machine-Translated versus Human-Localized Evaluation and Zero-Shot Transfer

A central goal of the benchmark is to measure performance under realistic multilingual deployment conditions. Beyond supervised evaluation on seen languages, we therefore examine two additional aspects of robustness: the gap between machine-translated and human-localized queries, and zero-shot transfer to unseen languages. Table 2 summarizes both analyses.

Machine-Translated versus Human-Localized Evaluation. Real customer utterances often contain shorthand, misspellings, colloquialisms, and other naturally occurring variation that translation pipelines tend to normalize. Under the controlled protocol in Section 3.2, machine-translated inputs yield higher accuracy in 42 of the 45 model-language settings, with a mean paired difference (machine-translated minus human-localized) of +1.319 accuracy points. A bootstrap over source-item IDs gives a 95% confidence interval of [0.159, 2.479]. The pattern is especially consistent for MMBERT, for which machine-translated evaluation is higher in all reported encoder conditions, by about 1.7 points on average. Representative exact McNemar tests on Chinese 17-way flat classification give $p = 0.0052$ for MMBERT and $p = 0.0192$ for Gemma 3 270M. Macro-F1 shows the same direction in 41 of 45 cells, with a mean paired difference of +1.819 points. Appendix D

gives the paired counts and procedures, and Appendix C gives the complete Macro-F1 results. We show one de-identified example. Both forms belong to the *Branch / Service Point Existence or Location* intent because they ask about the existence or location of a service branch. The machine-translated version is relatively literal, whereas the human-localized version is more conversational.

original: Isn't there a branch located south of the city?
machine-translated: 这座城市南部没有分店吗?
human-localized: 不是说在城市以南还有一家分店吗?
intent: Branch / Service Point Existence or Location

For Macro-F1, full results appear in Appendix C.

Zero-Shot Transfer to Unseen Languages.

Both columns of Table 2 are also zero-shot: models are trained only on English, Spanish, and Arabic and evaluated directly on Chinese, Indonesian, and Thai without target-language supervision. New markets typically open before localized labeled data exists, so this is the setting a deployed router starts in.

Model comparison. mmBERT slightly exceeds Gemma 3 270M on seen-language 13-way accuracy (94.69% vs. 94.08%), whereas Gemma 3 270M is higher in several Chinese and Thai zero-shot settings despite their similar sizes (270M vs. 307M). Gemma 3 1B has the highest result in most reported zero-shot settings.

Robustness to Granularity. Fine-grained intents are critical in production. Moving from 13-way parent to 17-way flat classification generally hurts BERT-style models more than Gemma models (Table 2), and causal LLMs naturally express hierarchical predictions whereas BERT-style classifiers require extra design. Thus, Gemma-based models fit settings emphasizing robustness to human-localized queries, finer taxonomies, and cross-lingual transfer, while BERT pipelines remain appropriate for conventional supervised classification. Appendix D reports limited checks on adaptation, an alternative translation engine, and concurrent throughput.

6 Conclusion

We have presented a public real-world benchmark for multilingual intent classification constructed

Table 2: Paired machine-translated and human-localized accuracy/Micro-F1 (%) on 674 shared source items per language; **bold** marks the best result within each model family, and N/A indicates that no parent-constrained decoder was implemented.

Model	Lang.	13-way Parent		17-way Flat		17-way Hier	
		MT	Human-loc.	MT	Human-loc.	MT	Human-loc.
mBERT	ZH	64.84	68.10	68.40	65.13	N/A	N/A
	ID	54.60	53.41	52.82	51.63	N/A	N/A
	TH	19.29	17.80	20.03	18.84	N/A	N/A
XLM-R	ZH	90.36	87.54	88.43	85.61	N/A	N/A
	ID	92.58	90.36	90.80	88.72	N/A	N/A
	TH	87.83	89.32	82.94	82.64	N/A	N/A
mmBERT	ZH	90.95	89.32	89.91	87.54	N/A	N/A
	ID	93.60	91.99	91.69	89.91	N/A	N/A
	TH	87.54	85.31	84.27	83.83	N/A	N/A
Gemma 3 270M	ZH	91.69	90.06	90.95	89.02	88.58	89.91
	ID	90.06	88.43	88.58	86.94	89.32	88.58
	TH	86.65	83.38	85.61	84.87	85.76	83.09
Qwen 3 0.6B	ZH	93.02	92.14	92.58	91.69	92.14	91.99
	ID	88.72	87.53	86.65	85.01	86.35	85.31
	TH	85.31	84.87	85.61	84.72	85.31	84.57
Gemma 3 1B	ZH	95.25	93.47	94.81	93.32	94.96	92.73
	ID	94.51	93.32	91.39	89.91	93.62	92.28
	TH	89.61	87.69	87.98	86.35	89.61	87.98

from logistics customer-service logs. It contains approximately 30K de-identified labeled training utterances organized into a partially hierarchical taxonomy with 13 parent and 17 leaf intents, covering three seen languages (English, Spanish, Arabic) and three test-only languages (Indonesian, Chinese, Thai) for zero-shot evaluation.

Across multilingual encoders and instruction-tuned small language models, scores are generally higher for machine-translated inputs than for professional localization (Table 2). The evaluated fine-tuned open models also reach higher fine-grained accuracy than the few-shot API references, though the two groups differ in both size and training regime.

We hope this benchmark helps evaluate multilingual intent systems under realistic long-tail distributions, noisy customer text, and cross-lingual transfer, and encourages professional localization alongside machine translation.

Limitations

Our benchmark is based on one logistics provider, one collection month, and six languages, so results may not generalize to other providers, domains, periods, or languages. The taxonomy is only partially hierarchical, and the rarest leaves have 7–16 examples per test-only language. The comparison between machine-translated and human-localized inputs uses filtered Spanish source items and one main machine-translation system, so the findings apply to this setting rather than every translation workflow. The paired interval does not capture fine-tuning-seed variability, and the alternative-engine check covers only two Chinese cells.

Ethical Considerations and Data Governance

Provenance, notice, and consent. The source interactions were collected in November 2025 from J&T Express operations in Mexico and Arabic-speaking markets. Under the platform’s user agreement and privacy policy, users were notified

that their interactions could be recorded and de-identified, and that the de-identified data might be used for model development and released for research. By using the service, users consented to these terms. The company’s legal and information-security teams also reviewed the intended use and release of the data.

De-identification and validation. Before annotation and release, we screened the data using deterministic rules and a privacy checker built on a locally deployed LLM. We then reviewed every retained utterance by hand and removed any example containing direct identifiers or combinations of contextual details that could create a re-identification risk.

This validation relied on manual privacy review; we did not conduct a quantitative evaluation of the de-identification pipeline. Because we did not construct an independently annotated PII test set, we do not report precision or recall for PII removal. We found no residual sensitive or identifying information during the final review. Because the data consist of free-form text, unexpected quasi-identifiers may still remain. We do not claim that the dataset is free of all privacy or re-identification risk.

Use, licensing, and redress. The dataset is released under CC BY-NC-SA 4.0. It is intended for research on intent classification and multilingual routing, not for identifying individuals, profiling customers, linking records to external personal data, or making automated decisions about individuals.

Anyone with privacy concerns or requests for correction or takedown can contact the corresponding author at the address listed above. If a concern is confirmed, we will correct or remove the affected examples from the canonical release.

Acknowledgments

We thank the annotation team and domain experts at J&T Express for their assistance with data curation, taxonomy design, and quality verification. We thank Bocheng Xia for their support and encouragement throughout this work. We also thank the anonymous reviewers for their constructive feedback. This work was supported in part by the J&T AI Group and ShanghaiTech University.

References

- Alibaba Cloud. 2025. [Qwen3-Max model information](#). Alibaba Cloud Model Studio Documentation.
- Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. [Revisiting machine translation for cross-lingual classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.
- Bytedance Seed. 2026. [Seed1.8 model card: Towards generalized real-world agency](#). *Preprint*, arXiv:2603.20633.
- 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 NLP for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.
- Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. [Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- DeepSeek-AI. 2025. [DeepSeek-V3.2: Pushing the frontier of open large language models](#). *arXiv preprint arXiv:2512.02556*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. [Towards debiasing translation artifacts](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. [TranslateGemma technical report](#). *arXiv preprint arXiv:2601.09012*.
- 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 Natara-jan. 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.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023. [NatCS: Eliciting natural customer support dialogues](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9652–9677, Toronto, Canada. Association for Computational Linguistics.
- Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. [An evaluation dataset for intent classification and out-of-scope prediction](#). 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 1311–1316, Hong Kong, China. Association for Computational Linguistics.
- Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. [MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2026. [mMBERT: A modern multilingual encoder with annealed language learning](#). In *Forty-third International Conference on Machine Learning*.
- Roman Plaud, Matthieu Labeau, Antoine Saillenfest, and Thomas Bonald. 2024. [Revisiting hierarchical text classification: Inference and metrics](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 231–242, Miami, FL, USA. Association for Computational Linguistics.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Pan-nyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathi-halli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Heseng Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. [EmbeddingGemma: Powerful and lightweight text representations](#). *Preprint*, arXiv:2509.20354.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.
- Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. [Cross-lingual transfer learning for multilingual task oriented dialog](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.
- Haoxin Wang, Xianhan Peng, Xucheng Huang, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Ling Jiang. 2025. [ECom-Bench: Can LLM agent resolve real-world E-commerce customer support issues?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 276–284, Suzhou (China). Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pi-eric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LLaMA-factory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset Supplement

Release metadata, split settings, and the hierarchy mapping behind Sec. 2 and Sec. 4.1 are collected here.

The 5% audit in Table 3 concerns training labels only. The seen-language test set was labeled independently by humans and adjudicated, and the test-only-language sets were constructed separately.

Table 4 provides the complete leaf-to-parent mapping used by the 13-way parent task and the 17-way flat/hierarchical tasks discussed in Sec. 5. Four parent categories contain multiple leaves because they cover operationally adjacent intents that are prone to confusion in customer-service routing: delivery change/intercept/return, damage-or-loss reporting/claims, service scope/transit/branch, and fee inquiry/dispute. The remaining parent categories map to a single leaf by design, because the corresponding business intents are already sufficiently separable and do not require artificial subdivision.

B Clustering Analysis

We adopted $K = 13$ for taxonomy construction on the following evidence.

To determine the optimal number of clusters, we conducted a systematic evaluation across $K \in \{10, \dots, 19\}$ using six complementary clustering quality metrics: Silhouette Score, Calinski–Harabasz (CH) Index, Davies–Bouldin (DB) Index, Gap Statistic, BIC, and Inertia. To enable direct comparison, all metrics were normalized to a common $[0, 1]$ scale and directionally aligned so that higher values consistently indicate better clustering quality; we then computed an aggregate score for each K . As illustrated in Fig. 3, $K = 13$ achieved the highest aggregate score (0.672) and offered the best overall trade-off across metrics (radar chart). In particular, it ranks among the top three configurations in terms of Silhouette (0.061), CH (1098.6), and DB (3.103; lower is better). The CH rebound from $K = 12$ to $K = 13$ is consistent with the additional cluster capturing useful structure. The elbow analysis agrees: the largest single-step drop in inertia falls at the $K = 12$ to $K = 13$ transition. $K = 13$ also yields the most balanced cluster-size distribution among all configurations, with a coefficient of variation of 0.392 and a minimum cluster size of 1,754.

C Macro-F1

The tables below give Macro-F1 under the same protocols and model groupings as Sec. 5.

Table 7 gives Macro-F1 for the machine-translated and human-localized conditions summarized in Sec. 5.

D Paired Analyses and Additional Checks

Paired inference. The observation that machine-translated accuracy is higher in 42 of 45 model–language–task cells is a descriptive consistency count, not 45 independent replications. On Chinese 17-way flat classification, mmBERT is correct on 606/674 machine-translated and 590/674 human-localized items; the discordant counts are 23 versus 7, giving a two-sided exact McNemar test of $p = 0.0052$. Gemma 3 270M is correct on 613/674 versus 600/674 items, with 20 versus 7 discordant pairs and $p = 0.0192$. For aggregate inference, a language-stratified paired cluster bootstrap resamples source-item IDs within each language and applies the sampled indices jointly to both conditions, all models, and all tasks. With 10,000 resamples and seed 30, the mean paired difference (machine-translated minus human-localized) is +1.319 accuracy points, with a 95% confidence interval of $[0.159, 2.479]$. This interval captures test-item sampling uncertainty, not fine-tuning-seed variability.

Macro-F1. Table 7 gives the complete paired Macro-F1 results. Machine-translated performance is higher in 41 of 45 cells, with a mean difference of +1.819 points; the mean gaps for the 13-way, 17-way flat, and 17-way hierarchical tasks are +1.638, +1.949, and +1.920 points. The four reversals are mBERT/ZH/13-way (−4.54), XLM-R/TH/13-way (−2.48), mmBERT/TH/17-flat (−0.53), and Gemma 3 270M/ZH/17-hierarchical (−0.36). Accuracy/Micro-F1 remains the primary operational metric, while Macro-F1 exposes long-tail behavior by weighting each intent equally.

Difficulty-floor baselines. On the translated 17-way set, majority-class, uniform-random, word/character TF-IDF with logistic regression, and subword fastText obtain 10.62%, 5.88%, 78.54%, and 64.69% Accuracy/Micro-F1, respectively. Under Thai zero-shot transfer, TF-IDF and fastText obtain 6.23% and 7.27%. These results show that head-class prediction does not explain

Table 3: Benchmark construction, evaluation, and release overview.

Field	Value
Raw source records	~600K multi-turn customer-service conversations from November 2025
Customer utterance units	~120K after removing agent turns and merging consecutive customer turns
Privacy and deduplication stage	~80K after rules, a local LLM privacy check, and exact/near/keyword duplicate removal
Training pool	~30K after quality, scope, and preliminary-label filtering
Seen benchmark languages	English, Spanish, Arabic
Test-only languages	Chinese, Indonesian, Thai
Seen-language test (Table 1)	~1.3K naturally occurring EN/ES/AR examples, all independently human-labeled with adjudication
Test-only-language evaluation (Table 2)	674 examples per condition and target language
Test-only class support	All 17 leaves per language; 7–93 items per leaf
Translation setup	Spanish source; google/translategemma-27b-it; local business translators
Training-label audit	Random 5% sample (~1,500); raw agreement ~95%; Cohen’s $\kappa \sim 0.94$
Release and licenses	Utterances, labels, language tags, taxonomy, fixed splits, paired test sets, code, prompts, and configurations at https://github.com/JTExpress-MICCS/MICCS ; dataset: CC BY-NC-SA 4.0; code, configurations, and prompts: CC0 1.0 Universal

the neural results and that the lexical baselines provide little cross-lingual transfer in this setting.

Alternative translation engine and adaptation.

On the same paired Chinese 17-way flat items, DeepL yields 90.06% for mmBERT and 90.95% for Gemma 3 270M, compared with 89.91% and 90.95% for TranslateGemma-27B. This limited two-cell check preserves the translated-over-human-localized direction but does not establish translation-engine invariance. As an adaptation control on the same seen-language split and Accuracy/Micro-F1 metric as Table 1, Gemma 3 27B increases from 79.46% with few-shot prompting to 97.16% after in-domain fine-tuning. This supports an effect of task-specific adaptation in this setting, not an inherent advantage of a particular model scale.

Concurrent throughput. On the same in-house concurrency benchmark using one NVIDIA RTX 3090, mmBERT-base processes 803 samples/s and Gemma 3 270M processes 414 samples/s. These measurements concern concurrent throughput only; they do not measure per-query latency, peak memory, monetary cost, or API serving performance.

E Training Parameters

Experiments are conducted with both encoder-based PLMs and decoder-only LLMs. For the encoder-based baselines, we use mBERT, XLM-R, and mmBERT. These models are fine-tuned using the Hugging Face Transformers library (Wolf et al., 2020) with a maximum input length of 64, a batch size of 128, a learning rate of 7×10^{-5} , weight decay of 0.01, and maximum gradient norm of 5.0. EmbeddingGemma uses the same classifier pipeline. For the decoder-only models, we evaluate Gemma 3 270M, Qwen 3 0.6B, and Gemma 3 1B, which are trained using LLaMA-Factory (Zheng et al., 2024). In LLM fine-tuning, we freeze all parameters except the last 16 transformer layers and use a batch size of 128 with a learning rate of 3×10^{-5} . All trainable models run for five epochs with AdamW, a cosine learning-rate schedule, and 150 warm-up steps. Each model receives nine hyperparameter configurations; the best is selected on the same seen-language development set. All experiments are conducted on a server equipped with 8 NVIDIA A800 GPUs.

Table 4: Classification label names and leaf-to-parent mapping for the released hierarchy.

Leaf ID	Leaf intent	Parent ID	Parent intent
0	Human Agent	0	Human Agent
1	Shipment Tracking / Delay	1	Shipment Tracking / Delay
2	Delivery Failure / Status Dispute	2	Delivery Failure / Status Dispute
3	Change Delivery Details / Method / Time	3	Delivery Change / Intercept / Return
4	Issue Report (No Compensation Request)	4	Damage / Missing / Confirmed Loss
5	Service Scope / Transit Time / Working Hours	5	Service Scope / Transit / Branch
6	Service Complaint	6	Service Complaint
7	Fee / Price Inquiry (No Dispute)	7	Fees / Price / Dispute
8	Pickup Booking	8	Pickup Booking
9	Shipment Details	9	Shipment Details
10	Packing / Restrictions / Insurance	10	Packing / Restrictions / Insurance
11	Invoice Request	11	Invoice Request
12	Business Partnership	12	Business Partnership
13	Intercept / Return Shipment	3	Delivery Change / Intercept / Return
14	Compensation / Claim / Claim Progress	4	Damage / Missing / Confirmed Loss
15	Branch / Service Point Existence or Location	5	Service Scope / Transit / Branch
16	Fee Dispute / Incorrect Charge	7	Fees / Price / Dispute

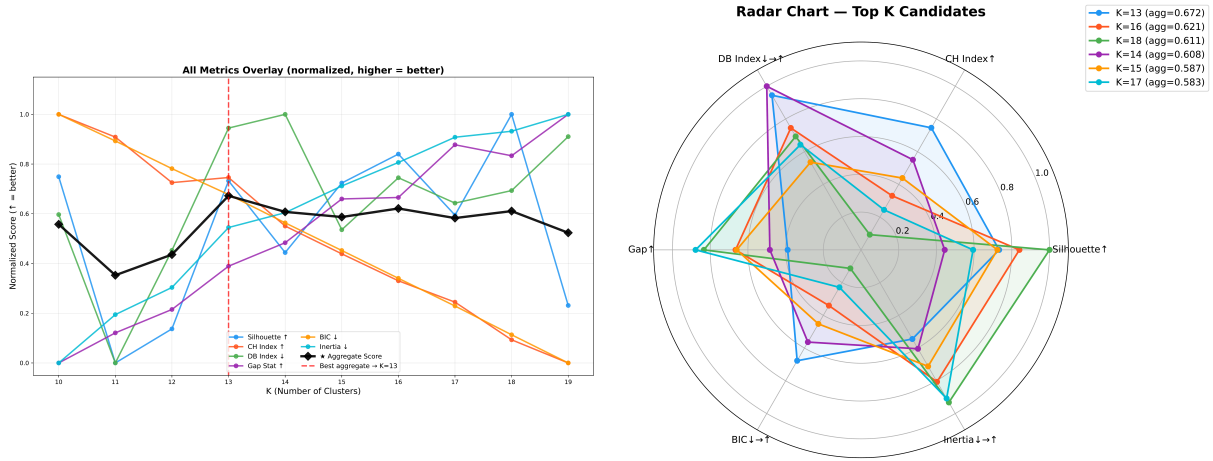


Figure 3: **Comprehensive evaluation of clustering configurations.** (Left) Normalized overlay of six clustering metrics across $K \in \{10, \dots, 19\}$. All metrics are scaled to $[0, 1]$ and direction-aligned so that higher values indicate better clustering quality. The black line represents the aggregate score, which peaks at $K = 13$ (indicated by the red dashed line). (Right) Radar chart comparing the multivariate performance profiles of the top-six K candidates. $K = 13$ achieves the highest aggregate score (0.672) and exhibits the most balanced, well-rounded performance across the evaluated dimensions.

Model IDs are jhu-clsp/mmBERT-base, google/embeddinggemma-300m, google/gemma-3-270m-it, google/gemma-3-1b-it, Qwen/Qwen3-0.6B, and google/translategemma-27b-it. API model IDs are doubao-seed-1-8-251228, deepseek-ai/DeepSeek-V3.2, and qwen3-max-preview. API decoding uses temperature 0; Appendix F shows the selected prompt format.

F System Prompt

Generation-based classification in Sec. 4.2 uses an XML-structured template with a role definition,

strict output rules, intent definitions, and few-shot demonstrations. Fig. 4 shows a representative excerpt.

Table 5: Companion Macro-F1 table for Table 1 on seen languages (%). Trainable open models use supervised fine-tuning; closed-source API LLMs use few-shot prompting only.

Model	13-way Macro	17-flat Macro	17-hier Macro
<i>Multilingual encoders</i>			
mBERT	90.85	88.75	N/A
XLM-R	92.86	90.15	N/A
mmBERT	93.82	92.10	N/A
<i>Embedding-based classifiers</i>			
EmbeddingGemma	94.12	92.35	N/A
Gemma 3 270M (embedding+classifier)	92.51	90.88	N/A
<i>Instruction-tuned SLMs</i>			
Gemma 3 270M	93.22	91.86	91.95
Qwen 3 0.6B	93.55	92.41	92.56
Gemma 3 1B	94.61	93.12	93.30
<i>Closed-source API LLMs (few-shot prompting; no fine-tuning)</i>			
Seed-1.8	92.74	92.57	91.31
DeepSeek V3.2	93.78	90.33	88.98
Qwen3-Max-Preview	95.41	91.48	93.64

Table 6: Parent-level Macro-F1 in 17-way hierarchical evaluation for models with hierarchical decoding (%).

Model	17-hier Parent Macro
<i>Instruction-tuned SLMs</i>	
Gemma 3 270M	93.22
Qwen 3 0.6B	93.55
Gemma 3 1B	94.61
<i>Closed-source API LLMs (few-shot prompting; no fine-tuning)</i>	
Seed-1.8	93.63
DeepSeek V3.2	92.51
Qwen3-Max-Preview	95.54

Table 7: Macro-F1 (%) for the machine-translated (MT) and human-localized conditions in Table 2.

Model	Lang.	13-way Parent		17-way Flat		17-way Hier	
		MT	Human-loc.	MT	Human-loc.	MT	Human-loc.
mBERT	ZH	57.35	61.89	61.89	59.93	N/A	N/A
	ID	53.56	52.93	51.78	50.64	N/A	N/A
	TH	17.62	15.66	18.04	17.93	N/A	N/A
XLM-R	ZH	87.13	82.29	84.74	80.99	N/A	N/A
	ID	90.73	88.85	88.14	85.96	N/A	N/A
	TH	85.68	88.16	80.67	80.19	N/A	N/A
mmBERT	ZH	90.57	87.20	87.89	84.54	N/A	N/A
	ID	90.66	88.86	88.37	85.85	N/A	N/A
	TH	84.54	84.02	82.14	82.67	N/A	N/A
Gemma 3 270M	ZH	90.85	88.25	89.75	86.82	87.05	87.41
	ID	88.56	85.93	86.78	84.14	87.32	85.58
	TH	84.15	79.88	82.81	81.07	82.76	79.09
Qwen 3 0.6B	ZH	92.42	90.64	91.78	89.89	91.14	89.99
	ID	87.22	85.33	84.85	82.51	84.35	82.51
	TH	83.31	82.07	83.41	81.72	82.81	81.37
Gemma 3 1B	ZH	94.85	92.27	94.21	91.92	94.26	91.13
	ID	93.71	91.82	90.39	88.11	92.52	90.28
	TH	88.11	85.49	86.18	83.85	87.61	85.18

Figure 4: Representative XML-style system prompt used for generation-based intent classification.

```
<system_prompt>
  <role>
    <title>Logistics Intent Classification Expert</title>
    <task>Analyze the user message and select exactly one intent ID.</task>
  </role>
  <core_rules description="Must be strictly observed">
    <rule id="1" name="Absolute Zero Misjudgment Principle">Only select an intent when the
      user's intent clearly fits the definition. If multiple intents appear, choose the
      dominant actionable one.</rule>
    <rule id="2" name="Output Format">Output ONLY the letter 'i' followed by the Integer ID
      (e.g., i7). No other words, punctuation, or explanations.</rule>
  </core_rules>
  <intent_classification_system>
    <intent id="0" name="Human Agent">
      <core_intent>User explicitly asks to talk to a real person.</core_intent>
      <definition>Request to transfer to human customer service.</definition>
      <examples>
        <example>Transfer to human customer service.</example>
        <example>I want to speak to a person.</example>
      </examples>
    </intent>
    ...
  </intent_classification_system>
  <few_shot_demonstrations>
    <demonstration>
      <user_input>My package hasn't moved for 4 days, can you urge them?</user_input>
      <output>i1</output>
    </demonstration>
    <demonstration>
      <user_input>Is it possible to ship a lighter and some fireworks?</user_input>
      <output>i10</output>
    </demonstration>
    <demonstration>
      <user_input>I want to speak to a person.</user_input>
      <output>i0</output>
    </demonstration>
  </few_shot_demonstrations>
</system_prompt>
```