

Spanish intent adaptation with LoRA

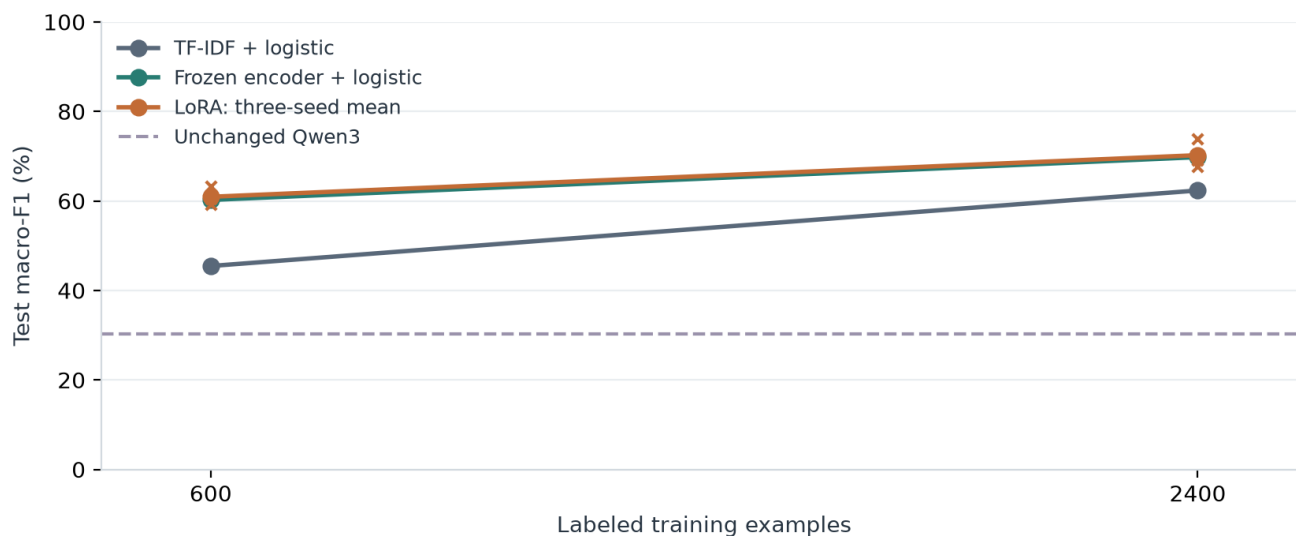
Qwen3-1.7B, two data budgets and smaller classifier baselines

Adapting Qwen3 changed test macro-F1 by +39.87 points (95% CI +37.72 to +41.67).

At 2,400 examples, LoRA / 2,400 / seed mean had the highest observed macro-F1 among the three fitted approaches. This comparison describes the fixed models and budget in this study; it is not a general ranking of the methods.

Final method	Macro-F1 %	Accuracy %	Invalid %
Qwen3 unchanged	30.35	30.87	4.51
TF-IDF / 600	45.45	54.81	0.00
Frozen encoder / 600	60.22	67.55	0.00
LoRA / 600 / seed mean	60.92	64.22	0.59
TF-IDF / 2,400	62.33	69.33	0.00
Frozen encoder / 2,400	69.78	74.82	0.00
LoRA / 2,400 / seed mean	70.22	75.26	0.20

Official test: 2,974 utterances / 60 intents. LoRA summaries average three fitted seed scores; they are not ensembles. All methods use the same test IDs and evaluator.



Crosses show individual LoRA seeds. The line connects two data budgets and does not establish a continuous scaling law.

A fixed comparison, with test kept separate

Data and smaller models

MASSIVE 1.1 es-ES supplies 11,514 official training rows, 2,033 development rows and 2,974 test rows. The same frozen nested samples of 600 and 2,400 rows train every fitted method. Only utterance text is a feature. Sampling seed: 20261008.

Word unigram/bigram TF-IDF and the frozen multilingual MiniLM encoder each feed logistic regression. Both choose C from {0.1, 1, 10} on development macro-F1, with accuracy and then lower C as tie breakers. MiniLM uses 384-dimensional, masked-mean, L2-normalized embeddings with a 128-token limit.

LoRA budget and selection

The four-bit Qwen3 base remains frozen. Rank-8 adapters cover all seven attention/MLP projections in the last four layers: 1,245,184 trainable parameters. Scale is 20, dropout is zero, and checkpointing is disabled. Adam uses a constant learning rate, betas (0.9, 0.999) and epsilon 1e-8.

Every fit makes exactly two passes. Batch four yields 300/1,200 optimizer updates at 600/2,400 examples. Loss averages real completion targets within each example and then across examples. Prompt and padded targets are excluded. No early stopping or checkpoint choice uses development results.

The development search includes three learning rates {1e-5, 3e-5, 1e-4}, three seeds and two sample sizes: 18 complete fits. Selection uses the three-seed mean development macro-F1; ties use mean accuracy then lower learning rate. All three seeds of each selected rate are retained.

Training rows	Selected learning rate	Selected seed IDs
600	0.0001	0, 1, 2
2,400	0.0001	0, 1, 2

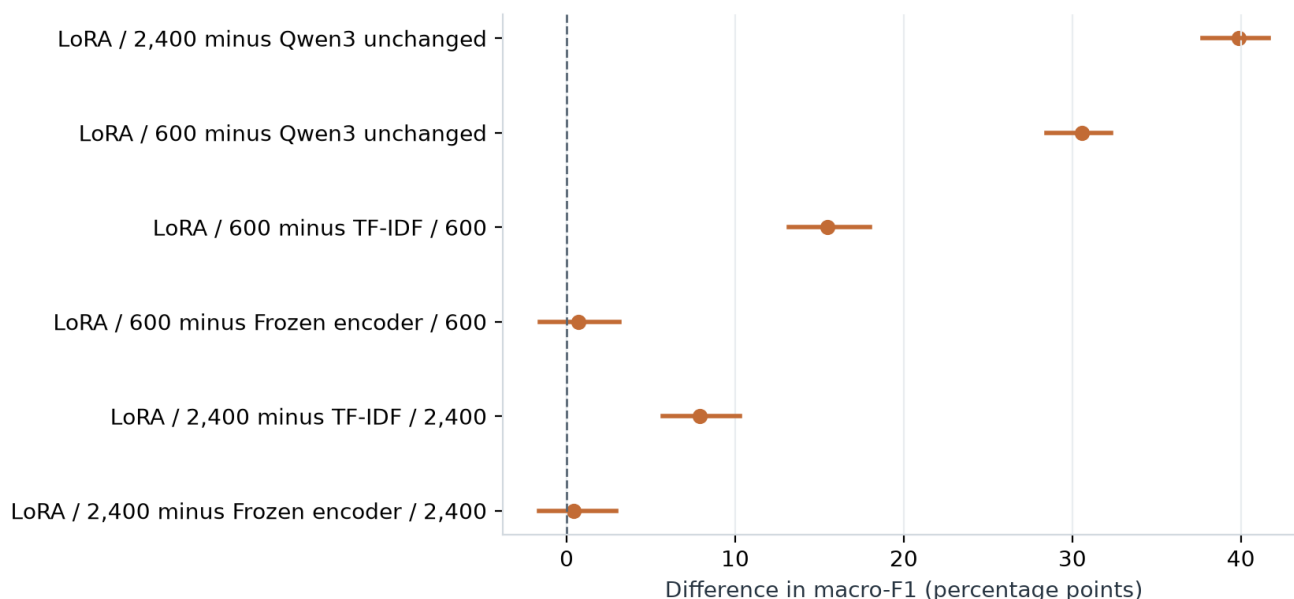
Common generation and scoring

Unchanged and adapted Qwen use the same Spanish instruction and alphabetical label list, no demonstrations, greedy decoding, thinking disabled and at most 32 new tokens. Only surrounding whitespace is removed. Explanations, wrong case and invalid labels count as errors. Macro-F1 retains all 60 intent classes; invalid output is not a 61st averaged class.

Quality evaluation uses canonical batches of eight. Batch-one engineering checks matched 16 sampled outputs before the study; that check is not a general equivalence guarantee. Interactive latency is measured separately at batch one.

The prospective comparative protocol followed known historical TF-IDF development scores and a five-update smoke run. It was not a preregistration before all data exploration. The joint freeze bound all 11 final instances before test scoring.

Paired uncertainty and overlap sensitivity



The primary contrast is LoRA at 2,400 rows minus unchanged Qwen in test macro-F1. Remaining contrasts are secondary. Intervals are 95% percentile intervals from 2,000 paired normalized-text-cluster bootstrap resamples (seed 2026). Every method receives the same sampled cluster multiplicities.

Intervals condition on the fitted adapters. They do not describe the full distribution of retraining outcomes. Secondary intervals are descriptive and have no multiple-comparison correction.

Training-seed variation

Budget	Seed 0 F1 %	Seed 1 F1 %	Seed 2 F1 %	Sample SD, points
600	59.16	63.20	60.41	2.07
2,400	73.79	69.12	67.75	3.17

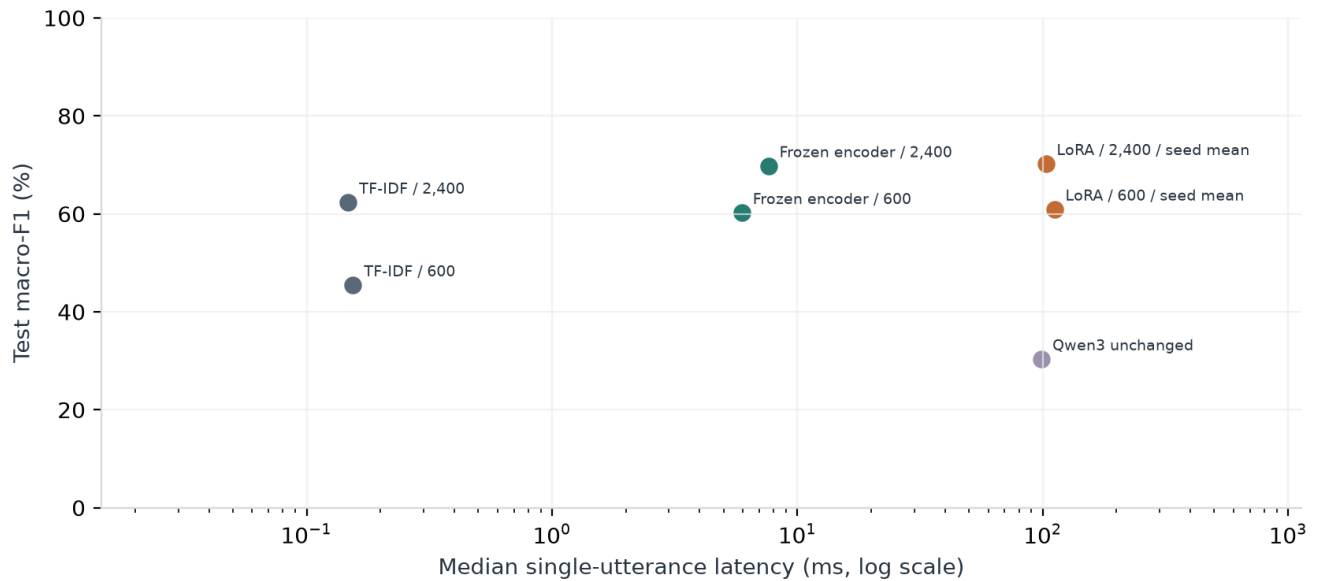
Prespecified overlap sensitivity

Remove test texts also present in the sampled 2,400 training rows or development set after NFKC, casefolding and whitespace collapse. This excludes 80 rows and leaves 2,894; every method uses the same retained IDs and all 60 labels remain in macro-F1.

Primary difference after exclusion: +39.70 points (95% CI +37.52 to +41.62).

The original full-data audit found 119 shared unique normalized text hashes between train/test and 39 between dev/test. Those are unique-hash counts for the official full splits, not affected-row counts for these samples. Exact-text filtering cannot address unknown paraphrase overlap or pretraining contamination.

Quality and local runtime costs



The same 100 development IDs are timed at batch one after one excluded warmup. Timing includes tokenization/features, inference and output selection. p95 uses nearest rank; loading is separate. LoRA points average three seed scores and three latency medians; individual timings follow. CPU and GPU pipelines have different backends.

Pipeline	Median ms	p95 ms	Load s
Frozen encoder / 2,400	7.63	370.96	5.75
Frozen encoder / 600	5.94	11.06	4.30
LoRA / 2,400 / seed 0	87.76	103.39	1.68
LoRA / 2,400 / seed 1	104.87	126.78	1.71
LoRA / 2,400 / seed 2	115.86	176.63	1.49
LoRA / 600 / seed 0	112.81	138.10	1.92
LoRA / 600 / seed 1	113.01	147.21	6.37
LoRA / 600 / seed 2	109.04	128.12	1.46
Qwen3 unchanged	98.85	112.53	1.57
TF-IDF / 2,400	0.15	0.21	0.61
TF-IDF / 600	0.15	0.24	0.44

All 18 LoRA fits: 98.02 minutes of training plus adapter save. Maximum training allocation: 3.264 GB (MLX); maximum process-lifetime RSS: 1.349 GB. These overlap and must not be summed.

Apple M5 Pro / 24 GiB unified memory. Shared-machine load and filesystem caches limit precision. Training-plus-save excludes development inference, initial loading and engineering work. Full timing records and per-fit exposure counts are preserved.

Interpretation, limits and reproducibility

What the result supports

Adapting Qwen3 changed test macro-F1 by +39.87 points (95% CI +37.72 to +41.67).

At 2,400 examples, LoRA / 2,400 / seed mean had the highest observed macro-F1 among the three fitted approaches. This comparison describes the fixed models and budget in this study; it is not a general ranking of the methods.

This is evidence about adapting a fixed small model for Spanish intent classification. It does not establish broader conversational quality, a new architecture or state-of-the-art performance. It is separate from production integrations such as Apple Foundation Models in Startup Journal.

Limits

The study covers one locale, one task and one small quantized LLM with a short fixed budget. Three seeds offer a limited view of training variability. Label frequencies are imbalanced. Models have unequal capacity and different pretraining; pretraining exposure to the evaluation material is unknown. The sensitivity addresses exact text overlap only. Resource measurements are operational observations from one shared machine.

Reproducibility record

The joint freeze records selected classifier weights, base model and tokenizer files, adapters, environments, protocol, training IDs, source hashes and all development evidence. Predictions are retained for every final method. A separate evaluator recomputes metrics, and the analysis cross-checks them independently to absolute tolerance 1e-12. Every search fit, including its training order and resource counters, remains available.

Freeze SHA-256:
d5d4f78fd4cbdb101907d5ee56df0c3102486cdcb893434743eb8fb6bd614857

The Markdown report and vector figures accompany this PDF. Machine-readable summaries contain per-intent precision, recall, F1, confusion counts, seed values, paired contrast intervals and the retained sensitivity IDs. See the adjacent analysis, evaluations, latency and source files for complete evidence.

Pinned model revisions

Qwen3-1.7B-4bit: 3b1b1768f8f8cf8351c712464f906e86c2b8269e
Multilingual MiniLM-L12-v2: e8f8c211226b894fcb81acc59f3b34ba3efd5f42

References

Hu et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models
FitzGerald et al. (2022). MASSIVE: A 1M-Example Multilingual NLU Dataset
Qwen3 four-bit MLX model card
Multilingual MiniLM model card
MLX-LM adaptation documentation