Shared-Probe Priors for Diagnosing and Guarding Against Expert-Routing Collapse in Multilingual ASR

Anonymous ACL submission

Abstract

Language-specific LoRA experts make multilingual ASR parameter-efficient, but they also turn language choice into a latent inference-time decision when labels are absent or unreliable. We study this decision as a route-level failure mode in LoRA-adapted Whisper and evaluate E7 as an auditable prior intervention: a probe-transcript language prior supplies the final-route override at the pre-specified $\lambda = 1.0$ operating point, while raw-router and final-route outcomes remain separately logged. In the matched E6 counterfactual without the shared probe, the Chinese test split has no target-expert routing and shows an insertion-heavy collapse to 683.33 CER; applying the E7 prior override recovers 99.87% Chinese target routing and reduces CER to 7.03. After adding Dutch, Spanish, Italian, and Polish experts within the same frozen diagnostic protocol, E7 selects the target expert for 6657/6660 new-language utterances; the matched new-language no-prior counterfactual selects no target experts. Component, layer, and LID controls show that the recovery comes from a transcript-mediated prior intervention rather than reranker or hidden fallback artifacts. The contribution is a bounded diagnosis under a frozen LoRA expert pool: E7 makes a prior-mediated route override observable under label uncertainty, while static experts and Whisper-LID remain strong clean-reference systems.

1 Introduction

Modern multilingual ASR systems must support existing languages while adding new ones without retraining a full model for each deployment. Whisper and large multilingual representation learners provide strong cross-lingual initialization (Radford et al., 2023; Babu et al., 2022; Zhang et al., 2023). Still, efficient adaptation remains delicate: full fine-tuning is costly, and adding languages can perturb existing behavior. LoRA reduces this cost by freezing the base model and learning low-rank updates (Hu et al., 2022). Recent ASR work uses shared LoRA experts or experts tied to language, speaker, and accent factors (Song et al., 2024; Li et al., 2025; Bagat et al., 2025). When language labels are absent or untrusted, however, the system must still infer which expert to use.

This selection step can fail more severely than the recognizer itself. In our matched counterfactual without a shared probe, Chinese utterances are routed to the wrong expert and produce an insertion-heavy CER collapse. Aggregate WER or CER alone cannot distinguish route choice, decoder looping, and expert quality. We therefore study expert selection as a latent decision and pair ASR error rates with route diagnostics: target-route distributions, prior sweeps, oracle gaps, and output-length checks. This follows a broader sparse-expert lesson: learned gates can develop collapse pathologies and often need stabilization (Shazeer et al., 2017; Chi et al., 2022; Fedus et al., 2022; Zoph et al., 2022; Zuo et al., 2022).

We study Shared-Probe Prior Intervention as a frozen diagnostic protocol for a fixed LoRA expert pool. E7 obtains an auxiliary shared-adapter probe transcript, estimates a language prior, and at the reported $\lambda = 1.0$ operating point uses this prior as the final selected-expert route while retaining raw encoder-router scores for audit. Unlike a clean LID gate, the protocol separates the failed raw route, the auditable prior intervention, and the final expert choice; the object of study is utterance-level routing collapse, not a new full-capacity MoE architecture.

Our key contributions are as follows:

- We identify expert-routing collapse as a distinct failure mode in LoRA expert inference without trusted language labels; in a matched counterfactual, Chinese test utterances never select the Chinese expert.

- We introduce a route-audit protocol that pairs WER and CER with target-route rates, expert distributions, oracle gaps, and output-length checks to separate routing failure from ordinary recognition error.

- We bound the shared-probe prior with retention and onboarding tests, prior sweeps, controls, and LID comparisons, reporting accuracy, latency, and auditability trade-offs.

The contribution is therefore a routing-collapse diagnosis and auditability study, not a leaderboard-style ASR comparison.

2 Related Work

2.1 Multilingual ASR

Whisper, XLS-R, and USM show the value of multilingual pretraining at scale (Radford et al., 2023; Babu et al., 2022; Zhang et al., 2023). Work on parameter-efficient extension asks how to add languages without full retraining (Liu et al., 2024). We focus on frozen Whisper-LoRA experts, where the central question is how to select language-specific updates when the language label is unavailable or untrusted.

2.2 LoRA Experts

LoRA adapts a frozen model with compact low-rank updates, avoiding full model copies (Hu et al., 2022). In ASR, LoRA-Whisper uses one adapter per language (Song et al., 2024); LoRA language experts and LoRA-MoLE fuse or distill monolingual experts for inference without a trusted language label (Li et al., 2025). HLoRA uses LID-posterior LoRA routing in a CTC ASR stack (Zheng et al., 2026), and mixture-of-LoRA work extends the idea to speaker or accent specialization (Zhao et al., 2024; Bagat et al., 2025). Layer-aware and rank-level LoRA variants study where sharing and specialization should occur during adaptation (Xu et al., 2026; Mei et al., 2026); our focus is routing-time diagnosis under label uncertainty.

2.3 Language Routing and MoE

Sparse MoE systems route each input to a small set of experts (Shazeer et al., 2017; Fedus et al., 2022), but representation collapse, training stability, and transfer stability remain central design issues (Chi et al., 2022; Zoph et al., 2022; Zuo et al., 2022). In ASR, LR-MoE uses shared LID routing to dispatch language-specific experts (Wang et al., 2023). BLR-MoE separates router confusion from ASR confusion through router augmentation and expert pruning (Ma et al., 2025), and Switch Conformer combines sparse language experts with universal phonetic experts (Mimura et al., 2025). Selective-invocation work shows that LID routing can add cost and suffer from misclassification (Xue et al., 2025).

2.4 Diagnostics and Evaluation

Layer-wise analyses show that acoustic and linguistic information changes with depth in speech representations (Pasad et al., 2021). Confidence and calibration diagnostics matter when ASR decisions depend on posterior scores (Woodward et al., 2020; Guo et al., 2017), and NLP significance guidance motivates paired tests with compact route-audit diagnostics (Dror et al., 2018). Recent work also highlights multilingual ASR metric pitfalls, beyond-WER Whisper diagnostics, and group-fairness disparities in multilingual ASR (Manohar and Pillai, 2024; Liang et al., 2025; Zee et al., 2024).

3 Setup and Route Audit

3.1 LoRA Expert Pool

The experiments use Whisper-small as the backbone. Language-specific LoRA experts use rank 32, alpha 64, dropout 0.05, and updates to `q_proj` and `v_proj` in encoder and decoder attention. Each adapter has 3,538,944 trainable parameters (3.54M in tables), verified from PEFT. The E7 shared-probe adapter uses the same counted unit. E4 uses one shared multilingual LoRA adapter, E5 uses static language-specific experts, and new-language onboarding inserts one added expert per new language.

3.2 Expert Selection

E6 and E7 build on the E5 experts by adding inference-time routing. Given an utterance, the system extracts encoder features, scores candidate experts, and decodes with the selected LoRA expert. E6 removes the shared-probe prior and serves as the matched no-prior counterfactual; E7 applies the prior intervention introduced below. Because E7 applies a text-based classifier to the probe transcript and fixes $\lambda = 1.0$, we describe it as an audit-path route override with an auxil-

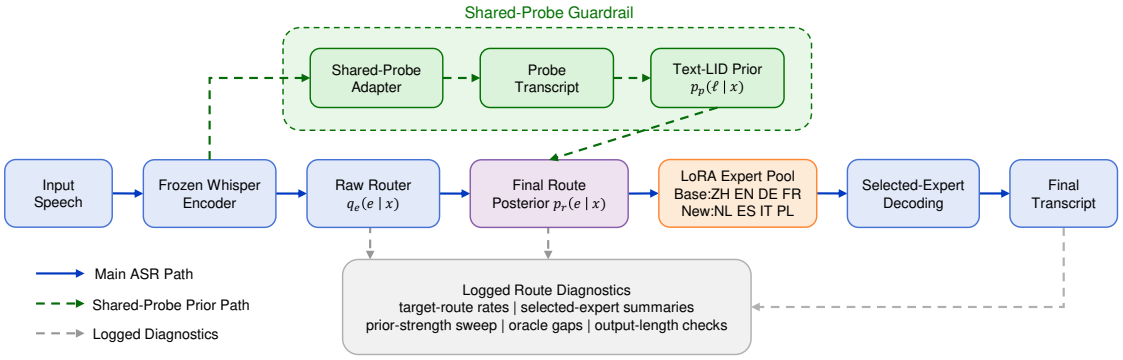

Figure 1: Shared-probe prior intervention and route diagnostics. The raw router remains separately logged, while the probe-derived prior supplies the final route at the audit operating point and the audit path logs route-level evidence.

iary transcript-LID prior. Unlike the Whisper-LID comparator, E7 exposes raw route, prior intervention, and final route traces; its value is diagnostic visibility under label uncertainty, not replacing explicit LID in clean accuracy or latency.

3.3 Collapse Diagnostics

The diagnostics use selected-expert summaries, target-route rates, oracle gaps, and output-length checks, with cached route traces for collapse reruns. We organize them around the E1–E7 ladder. E1–E4 are non-routed or shared-adapter references, E5 is the clean-label static-expert reference, E6 isolates the learned encoder router without the shared-probe prior, and E7 applies the prior while keeping the expert pool fixed.

We label a case as routing collapse only when route evidence and recognition symptoms agree: target routing is missing or badly degraded, one wrong expert dominates the selected-expert histogram, and oracle or length diagnostics show that a better route was available. If target routing is restored but error remains, we treat the residual as post-routing ASR headroom, not another collapse signal.

4 Shared-Probe Prior Intervention

4.1 Transcript Prior

The shared-probe path estimates the language evidence used by the reported route override before the final selected-expert decode. As shown in Figure 1, the raw router is trained from pooled Whisper encoder states from layers 6 and 8. We use these middle layers because speech cues vary with depth (Pasad et al., 2021) and the layer-pair control in Appendix Table A2 gives the best displayed Base Avg† for the tested layer-6 setting. E7 obtains a shared-adapter probe transcript and uses `langid.py` 1.1.6 (Lui and Baldwin, 2012) to form an auditable text-LID prior.

At the reported $\lambda = 1.0$ operating point, the shared probe supplies the deciding transcript-level evidence, preserving a clean boundary between raw router, probe prior, and final selected route.

4.2 Prior-Guided Selection

The encoder router remains logged for diagnosis, but at the reported $\lambda = 1.0$ operating point the final E7 route is a logged intervention by the probe-derived prior. Let $q_e(e \mid x)$ be the raw encoder-router posterior for expert e, $m(e)$ map expert e to its language, $p_p(\ell \mid x)$ be the probe-derived language posterior for language ℓ, and $\mathcal{E}$ be the expert set. E7 computes:

$$
\begin{aligned}
p_p(\ell \mid x) &= \mathrm{LangID}\big(\mathrm{ProbeASR}(x)\big), \\
q_e(e \mid x) &= \mathrm{softmax}_e\big(\mathrm{Router}(x)\big), \\
p_r(e \mid x) &\propto (1-\lambda)\, q_e(e \mid x) + \lambda\, p_p(m(e) \mid x).
\end{aligned}
$$

The final line is normalized over $e \in \mathcal{E}$. Reported E7 runs fix $\lambda = 1.0$ as a pre-specified audit point rather than a tuned interpolation: the final route is intentionally determined by the probe prior, while the raw router remains logged. Table 5 reports the frozen full-test sweep: weak prior fusion leaves Chinese routing unusable, whereas the collapsed route is avoided only once the probe prior dominates. Intermediate rows are diagnostics, not model-selection candidates; near-identical strong-prior rows indicate a stable override regime.

4.3 Controls

E6 and E7 use the same router-training recipe: expert supervision, class-balanced sampling, temperature scaling, and load-balancing regularization following standard sparse-expert practice (Zoph et al., 2022). The recipe is fixed across the comparison and is not an independent contribution. The

Role	Lang	Training sources	Train	Eval	Test
Base	ZH	AISHELL-1 150.9 + WenetSpeech S 100.1	250.9	18.1	10.0
	EN	LibriSpeech 100.6 + GigaSpeech S 250.0	350.6	5.4	5.4
	DE	MLS 150.0 + VoxPopuli 274.5	424.5	14.3	14.3
	FR	MLS 150.0 + VoxPopuli 215.6	365.6	10.1	10.1
New	NL	MLS 50.0 + VoxPopuli 46.4	96.4	12.8	12.8
	ES	MLS 49.9 + VoxPopuli 50.0	99.9	5.0	10.0
	IT	MLS 49.7 + VoxPopuli 50.0	99.7	5.0	2.5
	PL	MLS 50.6 + VoxPopuli 50.0	100.6	5.0	2.5

Table 1: Frozen data protocol after filtering. Durations are shown in hours and rounded to one decimal. Base languages support the collapse/diagnosis study; NL/ES/IT/PL are added only for the extension experiment.

controls rule out three alternative explanations for the observed recovery. First, the E6 and E7 boundary isolates the shared-probe prior while keeping the frozen experts and router recipe fixed. Second, prior sweeps, layer-pair checks, hard and soft routing checks, and no-reranker checks test whether the behavior comes from the logged prior intervention rather than a tuned endpoint, a special router layer, or hidden decoding fallback. Third, capacity, resource, and Whisper-LID controls bound the claim against stronger clean-label or explicit-LID references. Together, these controls support a narrow conclusion: E7 provides an auditable prior intervention for an observed routing collapse under the frozen protocol.

5 Experiments

5.1 Experimental Setup

Table 1 reports realized train, evaluation, and test hours for each language from frozen project splits. Chinese data come from AISHELL-1 (Bu et al., 2017) and WenetSpeech (Zhang et al., 2022); English from LibriSpeech (Panayotov et al., 2015) and GigaSpeech (Chen et al., 2021); German and French from MLS (Pratap et al., 2020) and VoxPopuli (Wang et al., 2021). The protocol fixes data manifests and sampling seeds before training. The main model runs use seed 42, as specified in Section 5.2. Sources differ in domain, speaker mix, and scale, so routing diagnostics are interpreted with WER and CER rather than as domain-independent behavior. For subset-based MLS and VoxPopuli data, speaker and duration distributions and source ratios are frozen before final evaluation. Main tables use only frozen test splits; development and pilot diagnostic runs are excluded. Leakage checks deduplicate splits by utterance ID, audio path, and speaker ID.

5.2 Implementation Details

All experiments used one NVIDIA A40 GPU. The stack is PyTorch 2.1, Hugging Face Transformers 4.36, PEFT 0.6, Datasets 2.16, and `langid.py` 1.1.6 (Lui and Baldwin, 2012). Main LoRA runs use effective batch size 32, learning rate 1e-4 with 500 warm-up steps and cosine decay to 1e-6, AdamW with weight decay 0.01, 20 epochs with patience-3 early stopping, rank 32, alpha 64, dropout 0.05, gradient clipping at 1.0, and seed 42 unless an ablation states otherwise.

For E7, the probe transcript is decoded with the shared LoRA adapter using one beam and automatic decode-language selection. `langid` is restricted to the candidate expert languages; at $\lambda = 1.0$, the final argmax is determined by the probe-derived prior, while the raw router posterior is logged only for audit. Decoder-side posterior-weighted language conditioning (weight 0.2) and confidence floors (0.60 base, 0.45 extension) were fixed before test decoding and are not the interpolation weights in Table 5. The independent frozen rerun that reproduced the E7 Base Avg† is reported only as a reproducibility check; all prediction records store raw-route probabilities, probe priors, final selected experts, and decoding hashes.

5.3 Systems

Table 2 defines the controlled base-language comparison. The rows separate system family, expert-selection rule, and stored-parameter cost: raw Whisper-small and full fine-tuning baselines, a shared multilingual LoRA, gold-language static LoRA experts (E5), Ours w/o shared-probe (E6), Ours with the shared-probe prior (E7), and a Whisper-LID gate that maps `detect_language` to an E5 expert. Under fixed data, expert pools, and decoding settings, the table tests whether routing decisions–rather than corpus or decoder changes–explain the E6/E7 gap. E6 stores the four E5 language experts plus the 9.2K encoder router; E7 adds one 3.54M shared-probe LoRA adapter, so its stored count is 3.54M$\times$5 + 9.2K rather than 3.54M$\times$4 + 9.2K. Resource, capacity, component, layer, and stress controls are reported in the appendix.

5.4 Metrics

Decoding is controlled separately from expert selection. All main comparisons use the recorded offline Whisper-small snapshot, beam size 5, fixed

ID	System	Selector	Stored Params	ZH	EN	DE	FR	Avg†
E1	Whisper-small raw	no	–	20.57	34.77	13.26	16.83	21.36
E2	Full-FT multilingual	shared model	240.6M	8.58	6.52	11.04	10.77	9.23
E3	Full-FT monolingual	gold language	240.6M×4	8.58	5.23	**10.81**	**10.10**	**8.68**
E4	Shared multilingual LoRA	shared adapter	3.54M	8.82	**4.67**	12.75	12.46	9.67
E5	Static LoRA experts	gold language	3.54M×4	**6.85**	5.55	11.34	11.40	8.79
E6‡	Ours w/o shared-probe	encoder router	3.54M×4 + 9.2K	683.33	5.54	33.91	11.49	183.57
E7	Ours: shared-probe prior	shared-probe prior	3.54M×5 + 9.2K	7.03	5.55	11.73	11.38	8.92
LID	Whisper-LID + E5 expert	Whisper-LID gate	3.54M×4	6.86	5.55	11.79	11.38	8.90

Table 2: Base-language clean-test results under the frozen protocol. ZH uses CER; EN/DE/FR use WER; Base Avg† is an unweighted descriptive average. Bold marks the best value among leaderboard/reference rows, excluding the E6 diagnostic row. ‡ marks the seed-42 full-test no-shared-probe collapse; full-test seed-43/44 checks are discussed in Section 8.

task setting, language-token policy, text normalization, generation temperature, and decoding configuration hashes stored with the result JSONs. Static experts, routed experts, and the Whisper-LID comparator therefore differ in adapter/expert selection, not in ASR decoding hyperparameters or oracle prompt information. All reported scores use frozen manifests and splits unless stated otherwise. ZH uses CER; all other languages use WER. Within a language, WER, CER, and oracle gaps are corpus-level micro averages over the displayed split. Base Avg† and New-language Avg are unweighted macro summaries across languages, not pooled cross-script ASR metrics. Paired bootstrap tests, where reported, resample aligned utterance-level predictions. Oracle gap is a route-audit margin, not an additional deployment setting: for each utterance, we compare the transcript produced by the selected expert with the lowest-error transcript among the same per-utterance candidate set of available experts. The reported Gap is the selected-route corpus error minus this per-utterance best-available corpus error, using the same metric scale as the table row (CER points for ZH and WER points otherwise). The target-language expert is therefore an operational proxy for the intended route rather than an absolute gold route, and we call collapse only when target routing is missed together with a large oracle gap or output-length pathology.

6 Results

6.1 Collapse and Prior Recovery

Table 2 reports clean base-language results including the E6 routing-collapse diagnostic. E6 shows an extreme Chinese failure and elevated German error, while the E7 prior path recovers recognition quality near strong clean references. E5 remains slightly better on Base Avg† (8.79 vs. 8.92), and Whisper-LID is also highly competitive (8.90).

Lang	Metric	E6: no probe			E7: probe			Diagnosis
		Err	Target	Gap	Err	Target	Gap	
ZH	CER	683.33	0.00	409.65	7.03	99.87	1.00	collapse→prior fix
EN	WER	5.54	100.00	2.76	5.55	99.96	2.77	stable
DE	WER	33.91	69.42	25.27	11.73	99.53	3.19	partial→prior fix
FR	WER	11.49	99.51	1.89	11.38	100.00	1.80	stable

Table 3: Route-level collapse diagnostics. Target is the percentage of utterances assigned to the target-language expert, and Gap is the selected-route corpus error minus the per-utterance best-available candidate corpus error in points; low E6 Target with a large Gap indicates routing failure, while high E7 Target with a small Gap indicates prior-mediated recovery.

Per-language paired tests show that E7 has higher error than E5 on ZH and DE, no reliable EN difference, and a small FR gain. Thus the clean-test result is best read as a successful prior-mediated route recovery near strong label-aware references, not as an accuracy win over static E5. We therefore do not make a formal non-inferiority claim against E5; the claim is prior-mediated route recovery with a small clean-label cost. An independent frozen rerun reproduced the E7 Base Avg† value.

Table 3 is included to make the mechanism claim explicit rather than to add another leaderboard. For each language, it asks whether E6's error is accompanied by missed target-expert routing and a large oracle gap, and whether the E7 prior path recovers target routing before the remaining recognition error is interpreted. This connects Table 2's aggregate scores to a route-level diagnosis.

Table 3 shows why the Chinese failure should be read as routing collapse rather than ordinary ASR degradation: the no-probe route misses the target expert and leaves a large oracle gap, while the shared-probe prior recovers target selection and leaves only small residual gaps. German shows a milder version of the same mechanism. Thus the table turns the aggregate error pattern into a route-level diagnosis: the main E6 failure is wrong expert selection, while residual E7 errors mainly reflect post-routing ASR limits.

6.2 New-Language Onboarding

Table 4 reports the eight-language onboarding view with both base- and new-language blocks. The base-language entries are rerun or retained references for the expanded onboarding protocol: E4 is retrained as the shared LoRA reference for the eight-language setting, E5 is the gold-language expert reference, and E7 reruns the expanded routed pool. Thus the block is a retention and protocol-alignment check rather than a replacement for the four-language leaderboard in Table 2. Under the expanded expert pool, E7 keeps Base Avg† at 8.92, indicating no measurable base-language degradation in this frozen protocol.

On new languages, E7 selects the target expert for 6657/6660 utterances. In WER, it is descriptively lower than E4 on all four added languages, reducing macro WER from 15.64 to 12.76 and nearly matching static experts (12.73). Paired-bootstrap checks show that the E7–E4 gains are reliable for IT and PL ($p < 0.001$) and small for NL and ES; the E7–E5 comparison is mostly tied except for a small Spanish gap favoring E5 ($p = 0.022$). Appendix Table A1 gives the matched no-prior counterfactual: without shared-probe, no added language selects its target expert and macro WER rises to 79.05. Raw Whisper-small remains strongest on NL and ES, so the claim is targeted routing stability and retention rather than uniform low-resource superiority.

7 Analysis and Ablations

This section keeps two reviewer-facing diagnostics in the body: the prior sweep and the route visualization. The appendix collects paired tests, new-language counterfactuals, layer and capacity checks, component checks, latency, and Whisper-LID stress results.

7.1 Prior Strength

Table 5 makes the prior boundary explicit on the frozen full-test split. The target-route block uses one column per language, so the override threshold is visible directly rather than through an ordered language list. Weak interpolation leaves Chinese routing unusable, whereas usable routing returns once the probe prior dominates. We keep $\lambda = 1.0$ as the named E7 point because it was the pre-specified, probe-prior-dominated operating point: after diagnosing raw-router collapse, the final route is chosen from the probe-derived prior rather than from a test-set-tuned mixture. The intermediate rows are frozen diagnostics, not model-selection candidates; the strong-prior rows only show that the override is not sensitive to the endpoint choice.

7.2 Probe Visualization

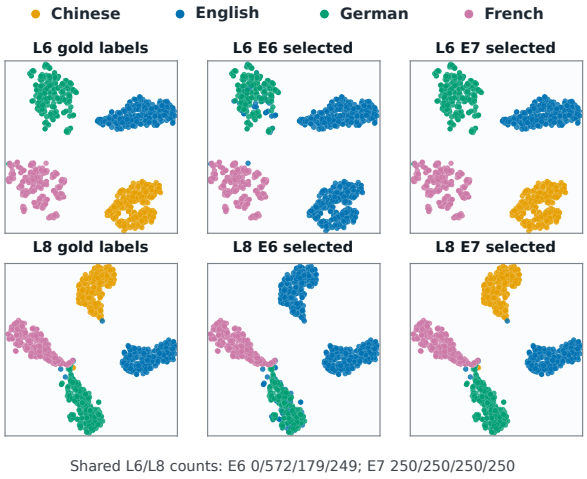

Figure 2: 2×3 t-SNE diagnostic over 250 samples per base language from Whisper layers 6 and 8. Columns show gold labels, E6 selected experts, and E7 selected experts; the shared count line reports selected-expert counts in legend order.

Figure 2 visualizes the shared-probe prior intervention as a qualitative route audit: gold-label panels show language structure, while E6/E7 panels contrast route decisions before and after the probe prior. It supports routing-collapse diagnosis and prior-mediated recovery, not perfect language separability.

7.3 Cost and Control Boundary

Appendix Table A3 quantifies the cost side of the trade-off. Whisper-LID is the stronger clean deployment reference when only accuracy and latency are considered: Table 2 gives Base Avg† 8.90 for Whisper-LID and 8.92 for E7, while Table A3 reports 904.3 versus 3538.2 ms/utterance. The distinction is diagnostic scope. Whisper-LID directly maps an utterance to an E5 expert, whereas E7 keeps the raw-router failure, probe-prior intervention, final route, and route-level diagnosis in the same auditable trace. We therefore do not present E7 as a faster or stronger LID gate; its role is to make the collapse and the prior-mediated repair observable under the frozen protocol.

8 Error Analysis

The error analysis separates three mechanisms that a single aggregate score would obscure. Table 3

ID	System	Selector	Stored Params	Base languages					New languages				
				ZH	EN	DE	FR	Avg†	NL	ES	IT	PL	Avg
E1	Whisper-small raw	no	–	20.57	34.77	13.26	16.83	21.36	**16.80**	**7.56**	17.36	17.77	14.87
E4	Shared multilingual LoRA	shared adapter	3.54M	7.60	5.83	14.21	12.54	10.05	17.70	8.04	17.66	19.15	15.64
E5	Static LoRA experts	gold language	3.54M×8	**6.85**	**5.55**	**11.34**	11.40	**8.79**	17.56	7.85	**11.76**	13.75	**12.73**
E7	Ours: shared-probe prior	shared-probe prior	3.54M×9 + 9.2K	7.02	**5.55**	11.73	**11.38**	8.92	17.61	7.94	11.78	**13.70**	12.76

Table 4: Eight-language onboarding results. ZH uses CER; all other language columns use WER. The E4 base-language cells are from the shared-LoRA rerun in the eight-language onboarding protocol, so they are not intended to duplicate the four-language E4 row in Table 2. E5 is the gold-language expert reference, and E7 is rerun with the expanded expert pool. Base Avg† is a descriptive mixed average over ZH CER and EN/DE/FR WER; New Avg is macro WER over NL/ES/IT/PL. Stored Params counts retained LoRA adapters plus the router; – marks no additional stored adapter/router.

Prior Setting	λ	Recognition error					Target route (%)			
		ZH	EN	DE	FR	Avg†	ZH	EN	DE	FR
Raw router	0.00	683.33	5.54	33.91	11.49	183.57	0.0	100.0	69.4	99.5
Fusion	0.25	683.33	5.54	13.58	11.47	178.48	0.0	100.0	96.1	99.6
Fusion	0.50	615.48	5.55	11.77	11.41	161.05	17.0	100.0	99.5	99.8
Fusion	0.75	7.03	5.55	11.75	11.38	8.93	99.9	100.0	99.5	100.0
E7	1.00	7.03	5.55	11.73	11.38	8.92	99.9	100.0	99.5	100.0

Table 5: Shared-probe prior sweep on the frozen full-test split. ZH uses CER; EN/DE/FR use WER. Splitting target-route rates by language shows the prior-strength boundary directly: Chinese remains collapsed through $\lambda = 0.50$ and enters the strong-prior recovery regime at $\lambda = 0.75$. E7 fixes $\lambda = 1.0$ as a probe-prior-dominated operating point rather than selecting a tuned interpolation.

Seed	N	Target (%)	Route hist.	H	CER	Gap
42	7176	0.00	EN:7176	0.5583	683.33	409.65
43	7176	0.00	DE:7176	0.5104	2096.19	1554.36
44	7176	0.00	EN:2649/FR:4527	0.5856	1693.67	1147.79

Table 6: Full-test E6 Chinese seed replication. Table 3 establishes the seed-42 route-level mechanism; this table tests whether the zero-target Chinese collapse persists across seeds. Target is the percentage routed to the Chinese expert; Route hist. is the selected wrong-expert histogram; Gap is the oracle CER gap in points. The replicated finding is zero target routing, while the wrong expert and CER magnitude are seed-dependent symptoms.

diagnosed the seed-42 route-level mechanism; Table 6 below asks whether the zero-target Chinese collapse is specific to that seed. The remaining subsections then separate residual post-routing error from language/source limits.

8.1 Routing Collapse

Table 6 shows that all three E6 Chinese runs assign 0.00% of 7176 utterances to the Chinese expert. The replicated fact is the absent target route; the dominant wrong expert and CER/gap magnitude are seed-dependent symptoms. This pattern is not consistent with a global label-map permutation: EN and FR remain stable in Table 3, E5 static experts decode the same manifests under trusted labels, and E7 restores Chinese target routing using the same expert indices. Seed-42 logs instead suggest a miscalibrated boundary: ZH is sent to EN despite $H = 0.558$, top-1 probability 0.752, top-2 margin 0.504, and coarse ZH posterior mass 0.018. This suggests, but does not prove, calibration pressure from class balance plus script/domain mismatch. The seed-42 Chinese trace is also insertion-heavy: the mean predicted/reference character-length ratio is 6.70, and 711/7176 utterances exceed a ratio of 6. After E7 applies the prior override, the ratio returns to 1.00 with no utterance above 6, so E6 is a controlled no-prior diagnostic rather than a competing ASR system.

8.2 Post-Routing Error

The E7–E5 gap is modest but systematic on some languages: +0.18 CER on ZH, -0.00 WER on EN, +0.40 WER on DE, and -0.02 WER on FR. The ZH and DE increases are reliable under paired bootstrap, EN is indistinguishable, and FR is slightly lower for E7. This is not coarse language-ID failure: E7 selects the German expert for 3378/3394 German utterances and the French expert for all 2426 French utterances. The remaining gap is therefore residual recognition cost after inferred expert selection; E5 remains the trusted-label upper baseline.

8.3 Language Limits

Stable expert selection is not uniform ASR improvement. Raw Whisper-small is strongest on Dutch and Spanish (16.80/7.56 WER versus E7 17.61/7.94), while Italian and Polish benefit more from expert onboarding (E7 11.78/13.70 versus raw 17.36/17.77 and E4 17.66/19.15). We therefore report per-language scores and retention summaries rather than claiming SOTA multilingual ASR or uniform gains.

9 Conclusion

This paper identifies expert-routing collapse as a distinct failure mode in LoRA expert ASR and provides a frozen diagnostic protocol for measuring it. E7 uses a pre-specified probe-transcript prior override to recover stable target routing in the observed collapse setting and after four-language expert expansion, while logging raw-router behavior, prior-sweep behavior, selected-expert traces, and oracle/length diagnostics. The evidence shows that the main failure is not weak expert capacity or ordinary ASR variance, but a latent expert-selection decision that can silently dominate recognition quality. By formulating the raw router, transcript prior, and final route as separate quantities, the method makes this decision observable under the evaluated protocol.

Limitations

The claim remains deliberately constrained. E7 is a prior-dominated audit path, not evidence that the raw learned router alone is reliable or that auxiliary LID evidence can be removed. Clean-label static experts and Whisper-LID remain strong references, new-language gains are uneven, and the audit path adds substantial latency (Table A3: 3538.2 ms/utterance versus 867.3 ms for E5 and 904.3 ms for Whisper-LID). The hard-vs-soft and no-reranker controls also show that aggregate improvements can come from less interpretable fallback behavior, so route transparency is part of the evaluation target rather than an optional diagnostic.

Future work should reduce probe cost through caching or selective invocation, learn confidence-aware prior weighting, and test learned audio-LID and matched-hour controls before extending the claim beyond the evaluated languages and corpora.

Ethical Considerations

This work uses public ASR corpora under their released research licenses and collects no new speech. AISHELL-1, WenetSpeech, LibriSpeech, GigaSpeech, MLS, and VoxPopuli are used only through frozen manifests and documented splits; no attempt is made to identify speakers. Per-language scores, routing histograms, stress behavior, and latency limits are reported because recognition quality can vary across languages, domains, speakers, and acoustic conditions. These diagnostics are not fairness guarantees: multilingual ASR can show group-level and intersectional performance disparities, so deployment requires additional group-aware auditing (Zee et al., 2024). The experiments are limited to Whisper-small adaptation and small LoRA experts, and we report parameter counts, storage requirements, and batch-1 latency so that readers can assess the method's resource footprint.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*, pages 2278–2282.

Raphaël Bagat, Irina Illina, and Emmanuel Vincent. 2025. Mixture of LoRA experts for low-resourced multi-accent automatic speech recognition. In *Proceedings of Interspeech*, pages 1143–1147.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *Proceedings of the 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment*, pages 1–5.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proceedings of Interspeech*, pages 3670–3674.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, pages 34600–34613.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1383–1392.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jiahong Li, Yiwen Shao, Jianheng Zhuo, Chenda Li, Liliang Tang, Dong Yu, and Yanmin Qian. 2025. Efficient multilingual ASR finetuning via LoRA language experts. In *Proceedings of Interspeech*, pages 1138–1142.

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. 2025. Beyond WER: Probing Whisper's sub-token decoder across diverse language resource levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31237–31247, Suzhou, China. Association for Computational Linguistics.

Wei Liu, Jingyong Hou, Dong Yang, Muyong Cao, and Tan Lee. 2024. A parameter-efficient language extension framework for multilingual ASR. In *Proceedings of Interspeech*, pages 3929–3933.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Guodong Ma, Wenxuan Wang, Lifeng Zhou, Yuting Yang, Yuke Li, and Binbin Du. 2025. BLR-MoE: Boosted language-routing mixture of experts for domain-robust multilingual E2E ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Kavya Manohar and Leena G. Pillai. 2024. What is lost in Normalization? Exploring pitfalls in multilingual ASR model evaluations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10864–10869, Miami, Florida, USA. Association for Computational Linguistics.

Yuxiang Mei, Delai Qiu, Shengping Liu, Jiaen Liang, and Yanhua Long. 2026. Zipper-LoRA: Dynamic parameter decoupling for speech-LLM based multilingual speech recognition. *arXiv preprint arXiv:2603.17558*.

Masato Mimura, Jaeyoung Lee, and Tatsuya Kawahara. 2025. Switch conformer with universal phonetic experts for multilingual ASR. In *Proceedings of Interspeech*, pages 1128–1132.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 914–921.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Proceedings of Interspeech*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-efficient and extensible multilingual ASR. In *Proceedings of Interspeech*, pages 3934–3938.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 993–1003.

Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. 2023. Language-routing mixture of experts for multilingual and code-switching speech recognition. In *Proceedings of Interspeech*, pages 1389–1393.

Alejandro Woodward, Clara Bonnín, Issey Masuda, David Varas, Elisenda Bou-Balust, and Juan Carlos Riveiro. 2020. Confidence measures in encoder-decoder models for speech recognition. In *Proceedings of Interspeech*, pages 611–615.

Jing Xu, Minglin Wu, Xueyuan Chen, Xixin Wu, and Helen Meng. 2026. Lamer-SSL: Layer-aware mixture of LoRA experts for continual multilingual expansion of self-supervised models without forgetting. *arXiv preprint arXiv:2602.12746*.

Hongfei Xue, Yufeng Tang, Jun Zhang, Xuelong Geng, and Lei Xie. 2025. Selective invocation for multilingual ASR: A cost-effective approach adapting to

speech recognition difficulty. In *Proceedings of Interspeech*, pages 2580–2584.

Anna Zee, Marc Zee, and Anders Søgaard. 2024. Group fairness in multilingual speech recognition models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2213–2226, Mexico City, Mexico. Association for Computational Linguistics.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6182–6186.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Qiuming Zhao, Guangzhi Sun, Chao Zhang, Mingxing Xu, and Thomas Fang Zheng. 2024. SAML: Speaker adaptive mixture of LoRA experts for end-to-end ASR. In *Proceedings of Interspeech*, pages 777–781.

Yuang Zheng, Yuxiang Mei, Dongxing Xu, Jie Chen, and Yanhua Long. 2026. A language-agnostic hierarchical LoRA-MoE architecture for CTC-based multilingual ASR. *arXiv preprint arXiv:2601.00557*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Jianfeng Gao, and Tuo Zhao. 2022. Taming sparsely activated transformer with stochastic experts. In *International Conference on Learning Representations*.

A Supplementary Checks

This appendix answers likely reviewer questions while keeping the main text focused on the routing-collapse story. Table A1 gives the matched no-prior counterfactual for the new languages, showing whether the new-language gains come from target-expert routing rather than expert quality alone. Table A2 compactly reports the key control evidence for likely artifact explanations: layer choice, reranker behavior, explicit LID substitution, expansion to eight experts, and stored capacity. Table A3 states the resource trade-off so the audit path is not mistaken for a low-latency deployment claim. Table A4 stress-tests the clean Whisper-LID reference under degraded audio, addressing whether an explicit LID gate makes the routing study unnecessary.

B New-Language Counterfactual

Table A1 isolates routing from expert quality for the added languages. It reports the matched no-prior counterfactual that the main text summarizes but does not need to tabulate in the body.

Lang	N	No prior	No-prior route	E7	E7 route
NL	3075	94.14	en:2982/de:93	17.61	nl:3074/en:1
ES	2385	56.22	en:1784/de:536/fr:65	7.94	es:2384/en:1
IT	594	53.40	en:441/fr:146/de:7	11.78	it:593/en:1
PL	606	112.46	en:549/de:57	13.70	pl:606
Macro	–	79.05	0 target	12.76	6657/6660 target

Table A1: New-language no-prior counterfactual. Without the shared-probe prior, no added language selects its target expert; E7 restores 6657/6660 target routes.

C Controls, Cost, and Stress

Table A2 summarizes the non-core controls by concern, check, and route-audit takeaway, so the appendix supports the method-boundary claims without expanding each check into a separate table.

Concern	Check	Takeaway
Layer pair	6/8=6/10 Avg† 8.92	Later pairs degrade ($\approx$9.9)
Hard gate	Avg† 8.93; min route 99.50%	Not a soft-routing artifact
No reranker	Avg† 8.71; default $\leq$21.80%	Fallback-heavy; audit routes
Explicit LID	Avg† 8.84; min route 99.50%	Strong clean reference
8-expert E7	Avg† 8.92; min route 99.53%	Base routing is preserved
Capacity	Rank-128 Avg† 9.37	Capacity alone insufficient

Table A2: Non-core controls on the frozen base-language split. Each row changes only the named factor and states the reviewer concern it addresses; Avg† follows Section 5.4.

Table A3 exists to make the cost boundary explicit: E7 is an audit-time prior intervention with added probe computation, not a speed or storage-efficiency claim.

System	Stored	Active params		Runtime	
		Fast	Audit	Fast	Audit
E2 full FT	240.6M	240.6M	240.6M	–	–
E4 shared LoRA	3.54M	3.54M	3.54M	–	–
E5 experts	14.16M	3.54M	3.54M	867.3	867.3
Whisper-LID	14.16M	3.54M	3.54M	904.3	904.3
E7 routed	17.70M+9.2K	3.54M+9.2K	7.08M+9.2K	2173.0	3538.2

Table A3: Resource and latency costs. Stored and Active are parameter counts; Runtime is batch-1 ms/utterance.

Table A4 qualifies the strong clean Whisper-LID reference under short/noisy speech. It is a

stress diagnostic, not a claim that E7 replaces explicit LID on clean accuracy or latency.

Stress	W-LID Avg	E7 Avg	W-LID acc.	E7 target route
short1s+SNR0	122.03	122.89	70.95	50.83
short2s+SNR10	85.69	84.99	95.36	91.06
short5s+SNR20	52.17	52.51	99.95	99.76

Table A4: Whisper-LID stress diagnostic under degraded audio. Scores use Base Avg†; W-LID acc. and E7 target route are macro percentages.