SLAM: A SELF-HOSTABLE MULTI-BACKEND RECONCILER FOR REAL-TIME MEDICAL ASR

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical automatic speech recognition (ASR) must recover safety-critical content while also applying formatting and editing operations to a clinical document. We study multi-backend ASR reconciliation, in which an LLM combines hypotheses from heterogeneous ASR systems into a single corrected and formatted output. We focus on a real-time per-chunk configuration that avoids the additional cost and state required by revisable rolling-window reconciliation. To make this configuration practical to deploy, we investigate replacing its proprietary frontier-model reconciler with a self-hostable open-weight student. We train students from the teacher's reconciliations using real clinician speech, synthetic speech, and their combination. Synthetic examples are language model generated, rendered with text-to-speech, and acoustically augmented to induce disagreement between the production ASR backends. We evaluate transcription and content fidelity on 226 held-out real clinical encounters and more complex formatting behavior on a separate 61-case targeted suite. Our best 27B student approaches the teacher, with no statistically significant difference in either word error rate (WER) or keyword error rate (KER). Synthetic supervision is competitive with real supervision at the evaluated training budget, while blended supervision matches or exceeds real-only supervision on the behavioral suite at every model size. Acoustic augmentation improves KER at every evaluated model size, with the largest observed gain for the smallest student. Self-hosting reduces measured reconciler cost by approximately 5x in our deployment comparison, and a 35B mixture-of-experts student reduces median reconciliation latency from 1,100,ms to 466,ms. These results show that the frontier-model reconciler in real-time multi-backend clinical ASR can be replaced by a faster, lower-cost, self-hostable model, with much of the required supervision generated synthetically.

1 INTRODUCTION

Ambient and voice-first interfaces are the fastest-growing surface for generative AI in clinical care (Tierney et al., 2024). Medical ASR must satisfy requirements for both clinical content accuracy and document behavior. The output must capture dense, critical content (drug names, dosages, patient identifiers, clinical terminology) that a single ASR model may mishear. Furthermore in clinical documentation settings, it must also produce the right document behavior by obeying the structural conventions of the destination field, turning spoken commands into formatting and preserving list structure across pauses. Raw ASR suffers from reliability here and the two requirements above make different demands on a system. We return to this distinction when we evaluate (Section 5).

SLAM addresses this with a post-processing stage that reconciles candidate transcripts from multiple ASR backends into a single corrected, formatted output. This combines the classical idea of recognizer-output voting across systems (Fiscus, 1997) with recent LLM-based generative error correction (Chen et al., 2023). Reconciling parallel raw hypotheses from heterogeneous backends, rather than trusting the output of any one backend, is the dominant source of quality in a clinical ASR pipeline (Makhervaks et al.). An LLM can fuse disagreeing hypotheses, ground them in structured patient context, and apply document-level formatting in a way that token-level voting cannot.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Multi-backend reconciliation admits a spectrum of configurations that trade recognition quality against latency and cost (Section 3). The production baseline uses a single batch-transcribed backend with LLM post-processing. Reconciliation occurs once a fixed window of speech is complete. Preliminary work by Makhervaks et al. extends this approach along two axes. First, it uses two ASR backends to transcribe the same audio independently and reconciles their outputs to improve recognition quality. Second, it performs reconciliation incrementally, allowing corrected text to be emitted after each speech chunk rather than after the full audio window. Two incremental configurations are considered: a per-chunk mini-batch ensemble (MBE), which processes each speech chunk once, and a rolling-window ensemble (RWE), which reprocesses recent audio and revises earlier output. Both fit a common formulation, where K is the number of chunks in each reconciliation window: K=1 for MBE and K>1 for RWE. Larger windows improve recognition quality but increase cost and implementation complexity. We therefore focus on MBE configuration with K=1.

Across these configurations, a proprietary frontier model performs reconciliation at each update, adding cost and latency to the real-time pipeline. We replace it with an open-weight small model that operates at near-parity with the frontier teacher and show that the supervision needed to train it can be manufactured synthetically. We evaluate the system with a comprehensive protocol spanning content fidelity, transcription, document behavior, and deployability. Our contributions are threefold:

- **A comparison of real-time reconciliation configurations in a clinical setting.** We formulate MBE and RWE along a common design axis and compare their recognition quality, cost, and implementation complexity. We identify the proprietary reconciler as the dominant contributor to per-encounter cost and latency, motivating its replacement with a self-hostable student in the MBE configuration (Section 3).
- **A distilled, self-hostable reconciler.** We replace the proprietary reconciler with a small open-weight model trained on the frontier model's own reconciliations. We characterize accuracy, latency, and cost. The best 27B student reaches near-parity with the teacher, while the 35B MoE student cuts median reconciliation latency by nearly 60% (Section 6).
- **Synthetic supervision with acoustic augmentation.** We construct a synthetic clinical speech corpus and show that synthetic supervision is competitive with real-speech supervision. Blending the two improves behavioral conformance, while acoustic augmentation improves KER across all evaluated model sizes (Section 6).

2 Related work

2.1 Recognizer combination and LLM-based correction

Combining multiple recognizers to reduce error is well established. ROVER aligns competing hypotheses and resolves them by voting, exploiting complementary recognizer errors (Fiscus, 1997). Voting, however, operates at the token level and cannot repair errors shared across systems or apply formatting. Recent work instead reframes correction generatively. Chen et al. (2023) generate a corrected transcript from an N-best list with an LLM, surpassing the oracle bound of reranking, a formulation since adopted by the GenSEC challenge (Yang et al., 2024). However, N-best hypotheses come from a single decoder and remain constrained by that recognizer. Ma et al. (2025) extend this across heterogeneous recognizers via a word-level confusion network. In a clinical setting, Makhervaks et al. reconciles raw parallel hypotheses directly, grounded in structured patient context, and introduces the two real-time configurations we compare. We build on it by establishing which configuration is deployable and by replacing its proprietary frontier reconciler with a self-hostable small model.

2.2 Synthetic speech, acoustic augmentation, and distillation

Clinical documentation consumes a substantial share of clinician time (Sinsky et al., 2016); clinical ASR can shorten time to note completion (Willoughby et al., 2025), and voice interfaces are among the fastest-growing applications of generative AI in care delivery (Tierney et al., 2024). Collecting labeled clinical audio at scale is nonetheless slow and privacy-constrained. Training ASR and downstream models on synthesized speech is established (Rosenberg et al., 2019), including in medicine, where LLM-generated data has hardened models against ASR errors (Binici et al., 2025).

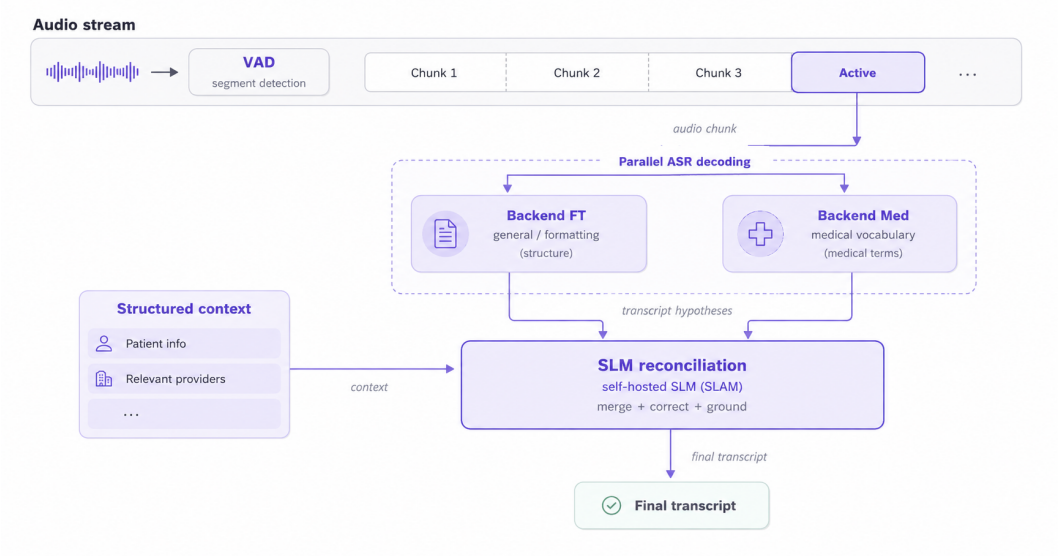

Figure 1: Overview of the SLAM inference pipeline.

Clean text-to-speech underrepresents real recognition noise, motivating acoustic and spectral augmentation from reverberant-speech training (Ko et al., 2017) to SpecAugment (Park et al., 2019); we adapt this to reconciliation by perturbing synthetic audio before replaying it through the production ensemble, inducing realistic backend disagreement to learn from. Our backends, an encoder-decoder transformer and a Conformer (Gulati et al., 2020), are representative modern recognizers; our contribution is finding the best reconciler configuration and making it self-hostable. We train open-weight students on the teacher's outputs via sequence-level knowledge distillation (Kim & Rush, 2016), which needs only generated text rather than teacher logits (Hinton et al., 2015), adapt them with LoRA (Hu et al., 2021), and study a MoE student (Jiang et al., 2024) for the accuracy-latency tradeoff under sparse activation.

3 Architecture and the real-time design space

3.1 System architecture

We adopt the architecture introduced in the preliminary work of Makhervaks et al.. A speech session is segmented by voice-activity detection (VAD) (Silero Team, 2024) into chunks. Each chunk is transcribed independently and in parallel by two ASR backends: Backend FT, a large multilingual encoder-decoder transformer pre-trained on weakly supervised audio and fine-tuned on clinical speech, and Backend Med, a 100M-parameter Conformer (Gulati et al., 2020) trained on de-identified clinical speech. Backend Med is strong on medical terminology and medication names but weak on proper names, a consequence of its de-identified training data.[1]

The reconciler receives both candidate transcripts, existing document context, and structured patient and encounter information, and produces a single corrected and formatted output. Figure 1 shows the inference pipeline. The experiments below focus on the mini-batch ensemble (MBE), as detailed below.

3.2 From design space to deployment

The production baseline applies LLM post-processing to a single batch-transcribed backend once the utterance is complete (B+PP). While substantially more accurate than raw ASR, it relies on a single backend and emits no text until the clinician stops speaking. The preliminary work addresses these

[1] We adopt the backend naming of Makhervaks et al.; the checkpoints themselves are proprietary.

Table 1: Cost and implementation trade-offs of incremental reconciliation.

	MBE (K=1)	RWE (K=3)
Revisable	No	Yes
Maintains state	No	Yes
Frontier cost (cents/session)	4.306	4.826
Qwen-4B cost (cents/session)	0.864	1.152

limitations by reconciling independently generated transcripts from two ASR backends, reducing normalized keyword error rate (nKER) by roughly 40% relative to B+PP, and processing speech incrementally through MBE or RWE.

We view MBE and RWE as points along a single design axis. At each update, the reconciler operates on transcripts generated from the most recent K VAD-aligned speech chunks. MBE is the K=1 case, producing incremental but non-revisable output. RWE uses K>1, allowing earlier output to be revised while its audio remains in scope. The RWE configuration evaluated here uses K=3. Full-batch and pure streaming form the extremes of the same axis (Appendix A).

Unlike B+PP, MBE and RWE emit text incrementally. Choosing between them requires balancing quality, latency, cost, and implementation complexity. MBE and RWE both provide incremental output, leaving cost and implementation complexity as the main practical differences. Self-hosting reduces reconciler cost by approximately 5x, motivating a fine-tuned small model in place of the frontier model. With Qwen-4B, RWE costs 33% more per session than MBE and requires maintaining state across overlapping windows. Although RWE retains a quality advantage (Appendix A), MBE provides the lower-cost and simpler operating point. We therefore focus on K=1 and ask whether fine-tuning a small, self-hosted reconciler can recover the quality of the frontier model.

4 RECONCILER: DISTILLING A FRONTIER MODEL ONTO A SELF-HOSTABLE SLM

The design choice at the center of this work is replacing the proprietary frontier reconciler with a self-hostable open-weight model. The student is trained on the frontier model's own reconciliations through sequence-level knowledge distillation (Kim & Rush, 2016; Hinton et al., 2015) rather than logit matching, since the proprietary teacher exposes only text. A key challenge is ensuring that the student generalizes across the clinical content and document behaviors encountered at deployment, requiring broad coverage in the fine-tuning data. Collecting real clinical speech at scale is costly and privacy-constrained. We therefore generate synthetic data and use the teacher's reconciliations as training targets. Beyond sidestepping that cost, synthetic generation has three properties we rely on. It bootstraps the training set, without waiting for production traffic to accumulate. It lets us deliberately construct the editing and formatting behaviors the reconciler must handle, such as empty-field composition, selection replacement, list handling, and spoken commands, directly and in balance. Finally, it covers the long tail of rare medications and unseen names (Appendix B.1).

Synthetic supervision. Figure 2 summarizes the pipeline. A language model generates clinical speech scripts spanning documentation contexts (notes, templates, and inbox messages), lengths, specialties, and delivery styles (pace, disfluencies, and run-ons). The scripts are modeled on plausible clinical encounters. Text-to-speech(TTS) renders the scripts to audio (Rosenberg et al., 2019; Binici et al., 2025), which we replay through the production ASR ensemble. This produces genuine backend errors and disagreements rather than simulating them in text. The closed-weight reconciler then generates the target labels. Since the resulting corpus contains no real clinical speech, evaluation on the held-out clinical set (Section 5) tests generalization from synthetic to real speech.

Acoustic augmentation increases the diversity of recognition errors in synthetic training data. In initial experiments, clean TTS audio produced similar transcripts across backends, providing limited examples from which the student could learn to resolve disagreements. Real speech produces exactly the discrepancies that create disagreement ("PERRL" → "Pearl", "DuoNeb" → "UNEP"). In turn, we added a post-TTS acoustic augmentation layer, in the spirit of acoustic and spectral

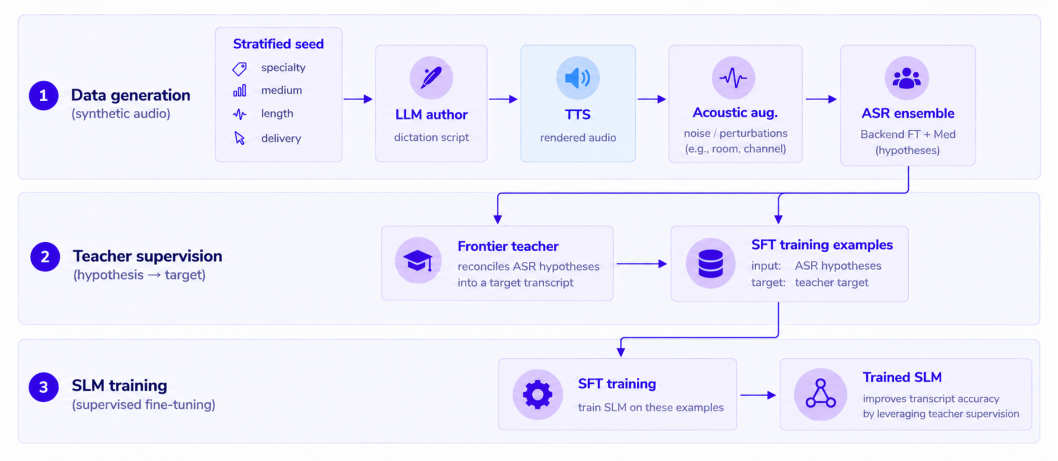

Figure 2: A clinical speech script is rendered by TTS and augmented to induce backend disagreement. ASR hypotheses and the frontier teacher's reconciled output form the input–target pairs used to train a small language model (SLM).

augmentation for robust ASR (Ko et al., 2017; Park et al., 2019). Each clip is perturbed by a sampled combination of additive babble, time compression, reverberation, onset clipping and band-limiting, all capped to a recoverable severity (see Appendix D for more details). The perturbations roughly double how far apart the two backends land while leaving the content recoverable. In the example below, augmentation truncates the drug name in one backend and drops the dosing frequency in the other, yet the teacher recovers both:

> **Backend FT:** "Number 2, continue glutide 1 milligram subcutaneous weekly. Number 3, maintain abdominal binder"
> **Backend Med:** "Number two, continue semaglutide 1 ml subcutaneously Number three to maintain abdominal binder."
> **Gold:** "Number 2, continue semaglutide 1 mg subcutaneously weekly. Number 3, maintain abdominal binder"

Backend FT preserves the dosing frequency ("weekly") but truncates the drug name, whereas Backend Med preserves the full drug name and route inflection ("subcutaneously") but drops the dosing frequency. The teacher combines the complementary information from the two backends, recovering both the full drug name and the dosing frequency. We evaluate the contribution of acoustic augmentation through a controlled ablation (Section 6.5).

5 Experiments

Datasets. We evaluate on two complementary held-out clinical test sets designed to measure distinct aspects of model performance. *ClinStandard* comprises 226 real clinical encounters spanning 28 specialties and 15 clinicians, with references produced through multi-pass expert annotation. Its size and breadth are intended to provide robust coverage of naturalistic clinical speech. *ClinBehavioral* is a separate, targeted suite of 61 cases designed to evaluate more complex note-editing behaviors that are not adequately characterized by standard encounter-level evaluation. These cases exercise specific clinician actions, including mid-note insertion, highlighted-selection replacement, and list handling across pauses. Rather than aiming for the same breadth as ClinStandard, ClinBehavioral is constructed to provide focused coverage of these distinct interaction patterns. Together, the two sets evaluate complementary dimensions of performance: broad performance on naturalistic clinical speech and targeted performance on clinically relevant editing behaviors.

Training data comes from three corpora: synthetic speech, real clinician speech, and their blend. Appendix B details their construction.

Comparative design. We fine-tune four Qwen3.5 model sizes (Team, 2026): 4B, 9B, 27B, and 35B-A3B. At each size, we evaluate the untuned base model and students trained with synthetic-only, real-only, and blended supervision, all against the proprietary frontier teacher. This evaluation design separates the effect of fine-tuning from model scale and isolates the contribution of synthetic supervision.

Metrics. We report KER, nKER, WER, and nWER on ClinStandard, the number of cases passed out of 61 on ClinBehavioral, and median reconciliation latency. KER measures content fidelity as the fraction of reference clinical keywords missing from the output, while WER measures overall transcription fidelity. Their normalized counterparts, nKER and nWER, remove casing and punctuation to separate content from formatting. Median (p50) per-call latency measures deployability on the real-time path. KER is the fraction of reference keyword occurrences missing from the hypothesis: for a unique keyword k with non-overlapping counts $c_{i,k}^{\text{ref}}$ and $c_{i,k}^{\text{hyp}}$,

$$\text{KER}_i = \frac{\sum_k \max\big(0,\ c_{i,k}^{\text{ref}} - c_{i,k}^{\text{hyp}}\big)}{\sum_k c_{i,k}^{\text{ref}}}$$

An LLM-based structured entity extractor identifies keywords in six categories: medications, medical terms, proper names, dates, numerical values, and locations.

WER is the standard word edit-distance rate,

$$\text{WER}_i = \frac{S_i + D_i + I_i}{N_i},$$

where S_i, D_i, and I_i are substitutions, deletions, and insertions from a Levenshtein word alignment and N_i is the reference word count. All rate metrics are micro-averaged over the corpus and reported raw and normalized; corpus figures are single trial (Section 5).

Experimental setup. We train LoRA adapters (Hu et al., 2021) (r=16) with learning rate 2×10^{-4} and linear decay, batch size 8, and maximum sequence length of 4096. For each configuration, we select the checkpoint on a held-out validation split and evaluate it on the test set. Student–teacher comparisons are paired at the encounter level across all 226 encounters. The 27B student was non-inferior to the teacher on KER and WER under pre-specified 10% relative margins (paired-bootstrap 90% CI upper bounds: $0.0056 < 0.00736$ and $0.0034 < 0.01336$ respectively).

6 Results and analysis

6.1 Supervision corpus and model size

Table 2 reports KER, nKER and ClinBehavioral pass rate for each model size trained on three corpora: fully synthetic clinical speech, real clinician speech, and their blend. Two findings stand out. First, synthetic supervision alone is competitive with real: on KER and nKER the two arms are close at every size, and neither consistently leads (Table 2). Second, the clearest advantage of blending is behavioral rather than in error rate. The blend passes 50–52 ClinBehavioral cases at every size, matching or exceeding real-only supervision at every size and strictly higher at three of four, the most consistent corpus effect in our study, whereas on KER and nKER no corpus consistently leads. Beyond the aggregate, the categorical gains from tuning land mostly in proper names, the hardest category for every base model (Section 6.4).

Fine-tuning brings every student close to the teacher. Relative to the untuned base of the same size, the blended student's KER falls (Table 3), landing within 14% of the teacher's KER and within 7% of its WER at every size (Table 2), including the single-GPU 4B. The 27B student trained on real clinician speech approaches the teacher's performance, landing within 0.003 KER and matching the teacher's WER, with no statistically significant difference on either metric.

Latency matters because the reconciler is invoked on every chunk on the real-time path. Every self-hosted student beats the proprietary foundation model's median of 1,100 ms, and the 35B-A3B MoE is fastest at 466 ms despite the most total parameters, since only $\sim$3B are active per token. The tightest-parity 27B and the fastest 35B-A3B are different points on this accuracy/latency frontier, so self-hosting lets a deployment pick the operating point it needs.

Table 2: Performance by model size and corpus. Latency is median per-call latency (p50).

Model	Supervision	WER	nWER	KER	nKER	ClinBeh.	p50 (ms)
4B	Synthetic	0.1469	0.0896	0.0910	0.0629	45	514
	Real	0.1398	0.0864	0.0889	0.0650	43	
	Blend	0.1419	0.0842	0.0838	0.0588	50	
9B	Synthetic	0.1446	0.0875	0.0871	0.0592	48	707
	Real	0.1370	0.0839	0.0858	0.0597	49	
	Blend	0.1374	**0.0809**	0.0814	0.0536	**52**	
27B	Synthetic	**0.1418**	**0.0821**	**0.0793**	**0.0504**	**49**	906
	Real	**0.1333**	**0.0816**	**0.0759**	**0.0506**	**50**	
	Blend	**0.1341**	0.0812	**0.0788**	**0.0522**	50	
35B-A3B	Synthetic	0.1451	0.0852	0.0799	0.0517	**49**	**466**
	Real	0.1365	0.0850	0.0835	0.0580	49	
	Blend	0.1420	0.0852	0.0819	0.0551	**52**	
Teacher	–	0.1336	0.0765	0.0736	0.0450	57	1,100

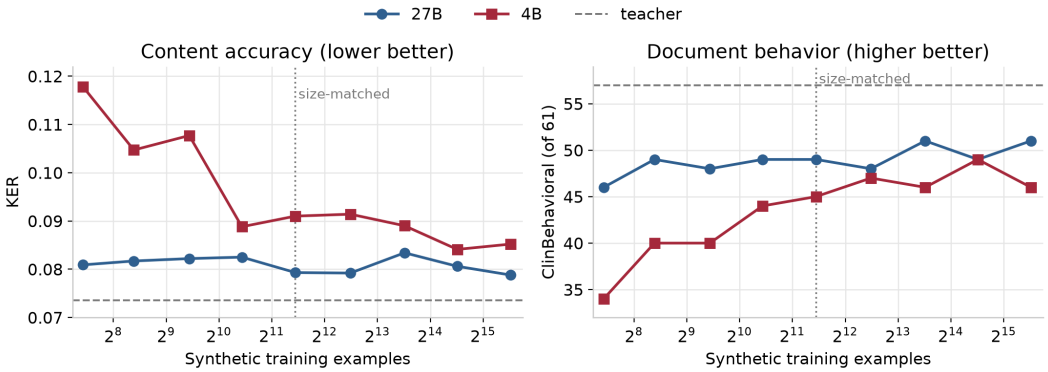

Figure 3: Synthetic dataset-size ablation with 4B and 27B students.

6.2 Scaling synthetic supervision

Given that synthetic supervision is competitive with real, we explore the effects of increasing the quantity of synthetic data. We sweep a nested synthetic-only corpus from 172 to 46,841 examples for the 4B and 27B students (Figure 3; per-tier values in Appendix C). The two model sizes scale differently. At 27B, just 172 examples capture nearly the full KER benefit, with KER varying by less than 6% as the dataset grows. Interestingly, behavioral performance continues to improve as the 27B gains five ClinBehavioral cases across the sweep. The 4B benefits more broadly from scale, reducing KER by 29% and gaining 15 ClinBehavioral cases. Thus, additional synthetic data has its largest effect on the smaller student, while still improving behavioral conformance at 27B after KER has saturated. The synthetic and blend experiments elsewhere use the 2,790-example tier, marked in Figure 3 and approximately matched in size to the real corpus.

6.3 The untuned-base model floor

In order to provide a proper frame of reference for how much our fine-tuning helps, we provide baseline metrics for each of the student models discussed. We evaluate each base model on ClinBehavioral and ClinStandard (Table 3).

There are two main findings that we observe. First, the 27B base is relatively strong out of the box (KER 0.0844), so fine-tuning buys only ≈7% on the 27B. However, the transformative gains are on the smaller models. Second, the base floor is not monotone in performance. The 9B base is anomalously weak, and the cause is instruction compliance rather than transcription: it frequently

Table 3: Untuned base vs. blended student performance at each model size.

Model		ClinBeh.	KER	ΔKER	KER by category				WER
					Med. terms	Numeric	Meds	Names	
4B	base	30	0.2386		0.2285	0.2249	0.2607	0.4028	0.3338
	blend	50	0.0838	−65%	0.0787	0.0782	0.1137	0.1627	0.1419
9B	base	32	0.3705		0.3711	0.3652	0.2784	0.4470	0.4173
	blend	**52**	0.0814	−78%	0.0781	0.0764	0.0991	**0.1360**	0.1374
27B	base	43	0.0844		0.0788	0.0905	0.0855	0.1596	0.1468
	blend	50	**0.0788**	−7%	**0.0749**	0.0802	**0.0753**	0.1459	**0.1341**
35B-A3B	base	38	0.1060		0.0996	0.0985	0.0983	0.2283	0.1902
	blend	**52**	0.0819	−23%	0.0783	**0.0762**	0.0894	0.1538	0.1420
Teacher		57	0.0736		0.0730	0.0674	0.0706	0.0955	0.1336

Table 4: Acoustic-augmentation ablation at the synthetic tier, **bold** marks the better arm.

Model		KER	nKER	WER	nWER
4B	aug	**0.0910**	**0.0629**	**0.1469**	**0.0896**
	no-aug	0.1002	0.0706	0.1596	0.1012
9B	aug	**0.0871**	**0.0592**	**0.1446**	**0.0875**
	no-aug	0.0878	0.0623	0.1479	0.0912
27B	aug	**0.0793**	**0.0504**	**0.1418**	**0.0821**
	no-aug	0.0827	0.0536	0.1421	0.0856
35B-A3B	aug	**0.0799**	**0.0517**	0.1451	0.0852
	no-aug	0.0819	0.0535	**0.1425**	**0.0849**

fails to emit the response at all, and omitted text is scored as deletion. Neither larger base shows this behavior. Part of the 9B's apparent fine-tuning gain is therefore a compliance gain, which is worth bearing in mind when reading its ΔKER (Appendix E).

6.4 Keyword error breakdown

In an effort to better understand where the reconciler's errors live, we decompose micro KER by keyword category and model size (Table 3), comparing each untuned base against its blended student. Across every size the story is consistent. Medical terms, 77% of the keyword mass, are near-saturated after tuning. The blend sits within 0.003 of the teacher at the 27B, so aggregate KER is dominated by a category tuning cannot move much. The blend's synthetic name coverage cuts sharply, most dramatically on the small models.

6.5 Acoustic augmentation ablation

Clean text-to-speech leaves the two backends almost nothing to disagree about, so a student trained on it never learns to reconcile. Degrading the audio before transcription is what supplies the signal. To separate that degradation from the effect of simply having more data, we train on the same synthetic tier with and without it, holding the row count fixed (Table 4).

Augmentation wins on both keyword metrics at every model size and on the word-level metrics everywhere except 35B-A3B. The backends disagree about drug names, acronyms and proper nouns, which is what KER measures and what degraded audio forces the student to resolve. The gain is largest at 4B and narrows as the student grows, since a larger model already recovers much of that disagreement unaided. Only the 4B and 27B keyword gains clear the training-seed variance for their cells, so the result here is the direction across sizes rather than any single difference. Appendix D reports the per-clip transforms, how far they move the backends apart, and the full set of deltas.

Table 5: Example reconciliations where the backends disagree.

Backend FT	"AO times 3, PERL, EOMI"
Backend Med	"O times three {comma} PERRL {comma} EOMI."
Reconciled	**"AO x3, PERRL, EOMI"**
Backend FT	"period, people's equal round reactive delight period."
Backend Med	"period} Pupils are equal, round, and reactive to light {period}"
Reconciled	**"Pupils are equal, round, and reactive to light."**
Backend FT	"General appearance, call in."
Backend Med	"General appearance, cocolon."
Reconciled	**"General appearance:"**

6.6 Qualitative examples

The behavior at the heart of the system is visible in individual reconciliations where the two backends disagree, on a clinical term, a garbled number, or a spoken command. The self-hosted reconciler produces the correct, formatted output rather than propagating either error (Table 5).

In the first example, the general-purpose backend drops a letter from 'PERRL', while the medical backend recovers the acronym but loses the leading 'A' from 'AO x3'. The reconciler combines the complementary evidence to recover "AO x3, PERRL, EOMI" and converts spoken commands to formatting. The third example is harder: neither backend recovers the spoken "colon," so the correct output cannot be obtained by token-level voting and must instead be inferred by the reconciler.

7 Discussion

Self-hosting and data residency. Self-hosting the reconciler allows candidate transcripts and clinical context to be processed within the operator's infrastructure rather than sent to an external model endpoint. This is important for data governance and simplifies deployment in settings where protected health information cannot cross an organizational boundary.

Limitations and future work. Three limitations most directly bound what we can claim. First, SLAM is trained on frontier-model reconciliations, which are themselves imperfect. A small set of clinician-authored targets could provide a stronger supervision signal. Second, our synthetic generation is not seeded on real speech. Conditioning the generator on patterns drawn from production audio could improve realism and narrow the synthetic-to-blend gap. Finally, we can extend to distill the RWE configuration ($K>1$) as it is a more complex architecture with the potential to yield improved performance.

8 Conclusion

We presented a real-time medical ASR system that replaces a proprietary frontier reconciler on the latency-critical path with a self-hostable open-weight model. The best 27B student reaches near-parity with its teacher, while a 35B MoE student runs at a fraction of the teacher's latency. Synthetic supervision is competitive with real clinician speech, and blending the two gives the most consistent gains on a behavioral test suite. Synthetic supervision also saturates with remarkably little data (Section 6.2), suggesting that further gains may come from better-targeted synthetic data blended with modest amounts of real speech rather than large clinical-audio collections. The synthetic pipeline can bootstrap supervision immediately and target rare medications, unseen names, and specific editing behaviors on demand, while the resulting model keeps audio and inference on operator-controlled infrastructure.

AI USE STATEMENT

In the creation of this paper, we used LLMs to polish our writing. This includes things like ensuring grammatical correctness and consistency. In addition, as part of the research process, we used LLMs to help retrieve relevant work and help us in the literature discovery process. Finally, as part of generating synthetic data in this work, we invoked LLMs as described in more detail in Section 4.

ETHICS STATEMENT

This work studies ASR for clinical settings. Transcription errors may signficantly impact patients and clinicians themselves. The systems evaluated here are intended to assist clinical documentation. Model outputs may contain errors and should not be treated as an authoritative source of clinical information without appropriate review and safeguards, which is part of our motivation for conducting this research.

Our experiments use both real clinician speech and synthetically generated clinical speech. Access to these data was restricted to the purposes and environments permitted under those controls. No actual PHI was used. We do not release any clinical data or the proprietary ASR checkpoints used in our production pipeline for these reasons.

REPRODUCIBILITY STATEMENT

Due to the proprietary nature of our work, we are not able to disclose the exact ASR models we use in our pipelines, however we do describe their architecture in Section 3.1. The fine-tuned reconcilers are all described to be part of the Qwen 3.5 family with their open-sourced base weights readily available. Details about the creation process and distribution of the training corpora can be found in detail in Appendix B. Full training hyperparameters, hardware, the evaluation pipeline, and known reproducibility caveats are provided in Appendix F.

REFERENCES

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F Chen, and Stefan Winkler. Medsage: Enhancing robustness of medical dialogue summarization to asr errors with llm-generated synthetic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23496–23504, 2025.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688, 2023.

Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE workshop on automatic speech recognition and understanding proceedings*, pp. 347–354. IEEE, 1997.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5220–5224. IEEE, 2017.

Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1389–1401, 2025.

Dave Makhervaks, Vinesh R Gudla, Kabir Mahal, Aditya Prabhakar, Josh Cowdy, Eric R Hunter, Raizy Leizerowski, Ryan Schwers, Madeline Grade, and Tejaswi Tenneti. Second thoughts: Revisable trailing context in llm-grounded clinical speech. In *Agentic AI for Scientific and Societal Advances*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pp. 996–1002. IEEE, 2019.

Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier, 2024. `https://github.com/snakers4/silero-vad`.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3):CAT–23, 2024.

Lauren R Willoughby, Daniel J Hekman, and Benjamin H Schnapp. The effect of dictation on emergency medicine resident time to note completion. *Western Journal of Emergency Medicine*, 26(6):1499, 2025.

Chao-Han Huck Yang, Taejin Park, Yuan Gong, Yuanchao Li, Zhehuai Chen, Yen-Ting Lin, Chen Chen, Yuchen Hu, Kunal Dhawan, Piotr Żelasko, et al. Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 371–378. IEEE, 2024.

A Real-time design-space evaluation

This section reports the quantitative comparison underlying the design choice in Section 3.2. Unless otherwise noted, results use the same 226-encounter held-out set and are micro-averaged.

Windowed formulation. Let the session audio be segmented by VAD into buffers $b_1, b_2, \ldots$. At tick t, the system processes the most recent K buffers,

$$W_t = (b_{t-K+1}, \ldots, b_t)$$

The two ASR backends produce hypotheses h_t^{FT} and h_t^{Med}, and the reconciler emits

$$y_t = R\left(h_t^{\mathrm{FT}}, h_t^{\mathrm{Med}}, c_{<t}, G\right), \tag{1}$$

where $c_{<t}$ is previously committed text and G is the structured patient/encounter context.

MBE corresponds to $K{=}1$: each buffer is reconciled once and its output is appended permanently.

RWE uses $K{>}1$: overlapping buffers remain in scope and their previously emitted text may be replaced. The standard RWE configuration is $K{=}3$. A buffer may therefore be transcribed and reconciled up to K times before leaving the active window.

MBE versus RWE. With the ASR backends, reconciler, and grounding held fixed, RWE reduces error relative to MBE across all four transcription metrics. Relative improvements are 22.3% for raw WER and 16.3% for raw KER, with corresponding improvements on the normalized metrics; all comparisons have $p < 10^{-3}$. Replacing the MBE prompt with the RWE prompt under a matched token budget does not recover this difference (Makhervaks et al.), indicating that the gain is associated with retaining and reprocessing recent audio rather than prompt wording alone.

Latency. The additional processing in RWE increases median per-tick latency from approximately 1,783,ms for MBE to 2,456,ms for RWE, a difference of about 0.7,s or 38%. The $p95$ latencies are similar. Because both configurations invoke the reconciler at every tick, reconciler inference is the primary target for reducing latency on the incremental path.

Multi-backend reconciliation. The larger quality effect is the use of multi-backend reconciliation itself. Relative to a single fine-tuned backend followed by LLM post-processing, combining two heterogeneous ASR hypotheses with structured grounding reduces nKER by approximately 40% (Makhervaks et al.).

B TRAINING CORPORA

Table 6 summarizes the supervision corpora used in the size-matched comparison. Synthetic examples are generated from language-model-authored clinical speech scripts, rendered with text-to-speech, optionally acoustically perturbed, and transcribed by the same ASR ensemble used at inference time. The proprietary teacher reconciliation is used as the training target. The real corpus consists of clinician speech collected through the production application and filtered to sessions containing outputs from both backends. Reference transcripts from the real corpus are not used as training targets.

Synthetic examples are generally longer than real examples. This difference in target length is a potential source of corpus-dependent behavior.

Table 6: Training corpora used in the corpus comparison (size-matched tier). The synthetic corpus is one tier of a nested ladder (172 to 46,841 rows) swept in Section 6.2.

Corpus	Rows	Composition
Synthetic	2,790	1,000 clips; 15.7% acoustically distorted
Real (staging)	2,965	production clinical speech
Blend	5,711	synthetic $\cup$ real, $\approx$51% real

Editing-context coverage. The synthetic corpus spans 27 clinical specialties, compared with 14 in the real training corpus. Table 7 compares editing context. Empty-field composition and the presence of text before and after the cursor are closely matched between corpora. Highlighted-selection replacement is more frequent in the synthetic corpus (45.0% versus 24.1%).

Table 7: Editing-context distribution, synthetic vs. real. Each row is the share of inputs with that context present (rows are not exclusive).

Context	Synthetic	Real
Empty field	31.3%	31.1%
Text before cursor	65.6%	66.5%
Text after cursor	54.7%	54.1%
Highlighted (replace)	45.0%	24.1%

B.1 Entity diversity and coverage

Clinical entities are sampled explicitly rather than generated freely by the language model. Patient and clinician names are drawn from a bank of 24,593 full names. Medications are drawn from a bank of 3,050 entities. The generation prompt requires assigned entities to appear verbatim. Under this sampling procedure, the most frequent first name accounts for 1.3–1.4% of name mentions across all corpus sizes.

Table 8: Entity coverage across the full ladder: distinct medications and distinct full names appearing in each tier. Tiers are named relative to the size-matched 1x point used in the main comparison (Table 2).

Tier	Rows	Medications	Names
$\frac{1}{16}$x	172	133	286
$\frac{1}{8}$x	335	269	601
$\frac{1}{4}$x	693	465	1,310
$\frac{1}{2}$x	1,383	738	2,499
1	2,790	1,185	4,461
2x	5,736	1,721	8,033
4x	11,686	2,258	12,932
8x	23,327	2,647	18,175
16x	46,841	2,918	21,642

C Synthetic scaling ablation

Table 9 reports the values underlying Figure 3. The tiers are nested with each tier containing all examples from the preceding tier.

Table 9: Synthetic-only ladder at 4B and 27B. Error rates are micro-averaged on *ClinStandard*; ClinBeh. is the pass count of 61. Tiers are nested, each a strict superset of the one below. The 1x tier is the size-matched point used for the synthetic arm elsewhere in the paper.

Model	Tier	Examples	KER	nKER	WER	nWER	ClinBeh.
4B	1/16x	172	0.1178	0.0872	0.1692	0.1081	34
	1/8x	335	0.1047	0.0721	0.1581	0.0935	40
	1/4x	693	0.1077	0.0775	0.1577	0.0964	40
	1/2x	1,383	0.0888	0.0594	0.1501	0.0872	44
	1x	2,790	0.0910	0.0629	0.1469	0.0896	45
	2x	5,736	0.0914	0.0612	0.1572	0.0985	47
	4x	11,686	0.0890	0.0662	0.1459	0.0933	46
	8x	23,327	0.0841	0.0584	0.1457	0.0886	49
	16x	46,841	0.0852	0.0583	0.1431	0.0866	46
27B	1/16x	172	0.0809	0.0545	0.1280	0.0731	46
	1/8x	335	0.0817	0.0553	0.1377	0.0814	49
	1/4x	693	0.0822	0.0545	0.1371	0.0823	48
	1/2x	1,383	0.0825	0.0547	0.1475	0.0862	49
	1x	2,790	0.0793	0.0504	0.1418	0.0821	49
	2x	5,736	0.0792	0.0543	0.1410	0.0861	48
	4x	11,686	0.0834	0.0546	0.1399	0.0855	51
	8x	23,327	0.0806	0.0552	0.1423	0.0883	49
	16x	46,841	0.0788	0.0519	0.1347	0.0799	51

D ACOUSTIC AUGMENTATION

Synthetic clips are optionally perturbed after text-to-speech synthesis and before ASR transcription. The augmentation family contains five transforms: multi-talker babble, time compression, reverberation, onset clipping, and telephone band-limiting. Table 10 lists the corresponding parameters.

Table 10: Per-clip acoustic transforms. Each is sampled independently per clip and applied to the text-to-speech output before transcription.

Transform	Parameter	Effect on the backends
Additive multi-talker babble	SNR (dB)	masks low-energy consonants
Time compression	rate	shortens phone durations
Reverberation	RT_{60} (s)	smears phone boundaries
Onset clipping	probability	removes utterance-initial energy
Telephone band-limiting	on/off	discards high-frequency detail

Augmentation frequency. Augmentation is applied to 20% of generated clips. Distorted audio clips produce fewer training rows on average, so 15.7% of rows in the size-matched corpus are augmented.

Inter-backend disagreement. Table 11 compares disagreement between the two ASR backends on clean and augmented examples. Mean normalized word-level disagreement increases from 0.19 to 0.48. The fraction of high-conflict chunks, defined as normalized word-level edit distance above 0.3, increases from 16% to 59%.

Table 11: Inter-backend disagreement on augmented vs. clean rows, over the full synthetic corpus. High-conflict chunks are those whose two hypotheses differ by a normalized word-level edit distance above 0.3.

	Clean	Augmented	Ratio
Rows	39,312	7,529	–
Mean word-level disagreement	0.19	0.48	2.6x
High-conflict chunks	16%	59%	3.7x
One backend omits the chunk	0.3%	13%	43x

E THE 9B BASE OUTLIER

The untuned 9B model has substantially higher error than the other base models. To determine whether this reflects transcription errors or incomplete outputs, Table 12 compares prediction length with reference length.

Table 12: Output volume of the untuned base models on ClinStandard. "Outputted" is total predicted words as a share of total reference words. "Median ratio" is the per-encounter prediction/reference word ratio. The next two columns are the share of encounters falling below 70% and 50% of their reference length. ρ is the per-encounter correlation between the fraction of reference words omitted and WER.

Model	Outputted	Median ratio	Enc. <70%	Enc. <50%	ρ(omitted, WER)
4B	91.4%	0.99	10.2%	1.9%	0.34
9B	71.0%	0.75	40.7%	22.0%	0.85
27B	100.8%	1.01	0.0%	0.0%	0.21
35B-A3B	102.6%	1.01	0.0%	0.0%	0.17

The 9B model emits 71.0% as many words as the reference corpus overall, compared with 91.4–102.6% for the other base models. Its prediction is shorter than 70% of the reference in 40.7% of encounters and shorter than 50% in 22.0%. Neither the 27B nor 35B-A3B base model produces an encounter below either threshold.

For the 9B model, the encounter-level correlation between the omitted-word fraction and WER is 0.85, compared with 0.17–0.34 for the other model sizes. These diagnostics indicate that incomplete output contributes substantially to the 9B base model's error. Its large improvement after fine-tuning should therefore be interpreted as a combination of improved output compliance and improved reconciliation.

F TRAINING AND EVALUATION DETAILS

This section records the training and evaluation configuration used for the reported experiments. Run manifests store resolved hyperparameters, dataset hashes, example counts, hardware presets, and per-epoch training and development losses.

F.1 MODELS AND ADAPTERS

We fine-tune Qwen 3.5 checkpoints (Team, 2026) at 4B, 9B, 27B, and 35B-A3B. The 35B-A3B model is a mixture-of-experts model with approximately 3B active parameters per token. All students use rank-16 LoRA (Hu et al., 2021). LoRA target modules and α use the managed training service defaults. No full-parameter fine-tuning is performed. Untuned baselines use the same checkpoints and evaluation pipeline without an adapter.

F.2 OPTIMIZATION

Table 13 reports the optimization settings used for all model sizes and supervision corpora. Sequences exceeding 4,096 tokens are excluded rather than truncated.

Table 13: Fine-tuning hyperparameters, identical across model sizes and corpora.

Setting	Value
Epochs	3
Learning rate	2×10^{-4}
LR schedule	linear decay to 0, no warmup
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8})
Weight decay	0
Gradient clip	1.0 (global norm)
Effective batch	8 sequences / optimizer step
Max sequence length	4,096 tokens

F.3 CHECKPOINT SELECTION

Each run is trained for three epochs. We select the epoch with the lowest loss on a held-out development set of 235 real examples, disjoint from the real training corpus at both the row and session level, and sharing no session or encounter with either evaluation set. Dev loss is used only to choose an epoch within a run. It is not comparable across supervision corpora, since the set is real audio and therefore in-distribution for the real and blend arms but out-of-distribution for synthetic-only.

F.4 HARDWARE

Training and evaluation use H100 GPUs with the allocations shown in Table 14.

Table 14: GPU presets per model, for training and for the evaluation sampler.

Model	Training	Inference
Qwen3.5-4B	H100x2	H100x1
Qwen3.5-9B	H100x2	H100x1
Qwen3.5-27B	H100x4	H100x4
Qwen3.5-35B-A3B	H100x8	H100x4

G THE CLINBEHAVIORAL SUITE

ClinBehavioral is a fixed suite of 61 cases. It is assembled partly from behaviors the reconciler must support by design and partly from real-life edge cases. (Table 15).

Table 15: Composition of the 61-case *ClinBehavioral* suite.

Category	Cases	Composition
Cursor and selection handling	32	mid-sentence insertion (12); long dictation replacing a selection (6); selection replacement (5); mid-document dictation with no selection (5); empty text-before with text-after (3); sentence-join at a selected boundary (1)
Transcription correction	8	medical term correction (4); homophone and contraction disambiguation (4)
Multi-chunk assembly	6	VAD-chunked dictation splitting across lines (3); plan and list reassembly (2); segment-boundary joining (1)
Mini-batch context discipline	6	behaviors with no full-batch analog (4); context not reproduced in output (2)
Spoken command conversion	5	spoken punctuation and formatting commands rendered as markup
Punctuation and field conventions	4	end-of-document terminal punctuation (3); template fields taking no terminal period (1)
Total	61	