How Small Can a Tandem Speech Front-End Be? Diagnosing Front-End Capacity with Layer Removal

Manato Yaguchi [*1,2] **So Kuroki** [*1]

Abstract

Recent tandem speech-to-speech (S2S) dialogue systems delegate knowledge and reasoning to a back-end LLM, leaving a front-end S2S transformer to handle low-latency spoken interaction. This raises a practical capacity question: how small can the front-end become, what performance is preserved, and how does its behavior change as layers are removed? To diagnose this, we randomly remove different numbers of layers from a KAME-style front-end transformer, fine-tune each variant on the same data, and evaluate on speech MT-Bench. The results show that moderate removal largely preserves performance, and even aggressive removal remains above the S2S-only Moshi baseline, although it no longer matches the full front-end. We then analyze individual dialogue turns to identify what fails. First, in low-scoring aggressive-removal cases, the model typically does not fall silent; it keeps answering, but the answers become long, incorrect, or unstable. Second, models with similar aggregate scores can still differ in performance at the level of individual dialogue turns. These diagnostic findings show why future front-end reduction should preserve both aggregate quality and dialogue-level behavior.

1. Introduction

Recent speech-to-speech (S2S) models integrate speech perception, language modeling, and speech generation into a single interactive model (Défossez et al., 2024; Zhang et al., 2023), with newer omni-style systems extending this direction to broader multimodal and full-duplex settings (Xie & Wu, 2024; Wang et al., 2025; Xu et al., 2025). This consolidation reduces cascade overhead, but leaves real-time models to manage multiple interaction functions at once.

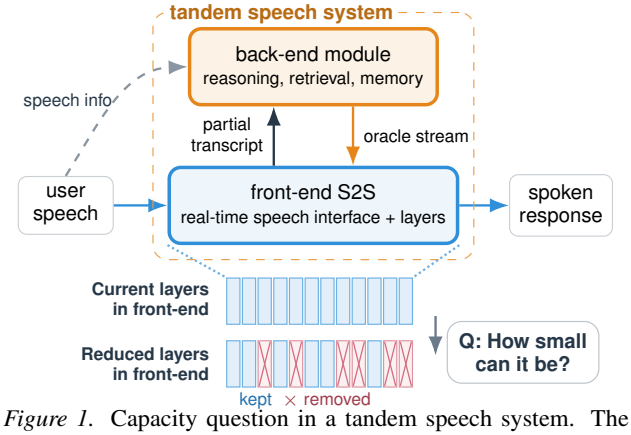

Figure 1. Capacity question in a tandem speech system. The front-end stays in the real-time speech path, while the back-end supplies semantic guidance through the oracle stream. Layer removal probes how much front-end depth must remain for this interface to keep working after recovery fine-tuning.

Tandem speech systems address this burden by separating the real-time speech interface from slower knowledge and reasoning modules. KAME keeps a front-end transformer in the low-latency speech path while a back-end LLM supplies evolving candidate responses through an oracle stream (Kuroki et al., 2026). Related systems similarly explore asynchronous knowledge access for real-time speech dialogue (Chien et al., 2026). This separation makes the capacity question in Figure 1 natural: *if knowledge-intensive behavior can be delegated to the back-end, how small can the front-end transformer become?*

Existing layer-removal and compression studies do not directly answer this question. They show that transformer depth can be redundant in language and speech models (Fan et al., 2020; Men et al., 2025; Kim et al., 2024; Zhong et al., 2025; Dorszewski et al., 2025), but reducing S2S dialogue models remains less explored. Moreover, a tandem front-end has a different role from a standalone language model or speech encoder: it must coordinate streaming speech, back-end guidance, multi-stream state, and speech realization, so layer removal may change its dialogue behavior. Thus, the front-end capacity question remains open in this setting.

[*]Equal contribution [1]Sakana AI, Tokyo, Japan [2]The University of Tokyo, Tokyo, Japan. Correspondence to: Manato Yaguchi <manatoyaguchi@sakana.ai>, So Kuroki <sokuroki@sakana.ai>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

To diagnose this question, we use budgeted random layer removal on the front-end transformer. For each removal budget, we randomly choose which layers to remove, keep the back-end and stream interfaces fixed, fine-tune each variant with the same oracle-conditioned data, and evaluate on speech MT-Bench. In this setting, random removal is not a rule for selecting the best layer combination. Instead, we treat it as a perturbation for asking how well a reduced front-end can recover, where that recovery breaks, and what kinds of failures emerge.

The results first show a macro-level capacity pattern: the KAME-style front-end has recoverable but bounded depth slack, with moderate removal largely preserving performance and occasionally matching or exceeding the full front-end, while aggressive removal produces a clear cliff. We then look below aggregate scores and find two dialogue-level effects. First, the cliff is not simply a failure to answer; low-scoring aggressive-removal cases often continue speaking, but become long, incorrect, or unstable. Second, variants with similar overall scores can still differ across individual dialogue turns. Together, these findings suggest that front-end reduction should be studied as both a capacity problem and a behavioral problem: how much can be removed, where failure begins, and which dialogue behaviors remain reliable.

2. Layer-Removal Diagnostic Protocol

This section defines a controlled layer-removal diagnostic for a KAME-style tandem front-end (Kuroki et al., 2026). We first specify the front-end under study, then define budgeted layer removal, recovery fine-tuning, and evaluation while the surrounding tandem system is held fixed.

2.1. KAME Front-End Under Study

A KAME-style tandem system keeps a real-time S2S front-end in the speech path while a back-end LLM supplies an oracle stream. This front-end receives the streaming speech representation, conditions on the oracle stream, maintains the inner-monologue stream, and generates the spoken response. We therefore intervene only on the transformer stack inside this front-end; the back-end model, oracle construction, audio modules, stream interfaces, tokenizers, output heads, decoding configuration, recovery fine-tuning, and evaluation protocol are shared across conditions.

Let the front-end transformer contain $L = 32$ layers, denoted by $\{f_1, \ldots, f_L\}$. A reduced front-end is specified by a binary mask $m \in \{0,1\}^L$, where $m_\ell = 1$ retains layer ℓ and $m_\ell = 0$ removes it. For removal budget K, a valid mask satisfies

$$\sum_{\ell=1}^{L} (1 - m_\ell) = K. \tag{1}$$

We protect the first $C = 3$ transformer layers to avoid perturbing the lowest-level stream interface.

For an input hidden state h_0, the masked forward pass is written over the original layer positions as

$$h_\ell = \begin{cases} f_\ell(h_{\ell-1}), & m_\ell = 1, \\ h_{\ell-1}, & m_\ell = 0, \end{cases} \quad \ell = 1, \ldots, L. \tag{2}$$

Retained layers are copied from the full front-end and applied in their original order, while removed layers are bypassed by the identity map. Since all stream-level input and output dimensions are unchanged, every reduced model remains compatible with the same oracle-conditioned training and inference pipeline.

2.2. Layer-Removal Diagnostic Procedure

Step 1: sample removal masks. For each budget K, we sample three independent valid masks uniformly over the unprotected transformer layers. The main diagnostic budgets are $K \in \{3, 5, 7, 10\}$, covering small, moderate, and aggressive reduction regimes. These sampled masks are diagnostic perturbations, not random pruning baselines: their role is to estimate depth slack and mask sensitivity under a fixed removal budget.

Step 2: recovery fine-tuning. After applying a sampled mask, we fine-tune the reduced front-end with the same oracle-conditioned multi-stream SFT objective used for the full tandem system. All variants use the same data distribution, oracle construction, back-end responses, and decoding setup. All reported models, including the $K = 0$ full front-end, are evaluated under the corresponding 3000-step oracle-conditioned SFT setup. Thus, the diagnostic asks whether the tandem interface can absorb a given removal budget after a fixed amount of adaptation, rather than whether the layer-skipped model works immediately after surgery.

Step 3: evaluate and aggregate. We evaluate each model with the speech-synthesized MT-Bench protocol used for KAME (Kuroki et al., 2026), following the LLM-as-a-judge MT-Bench framework (Zheng et al., 2023). We focus on three knowledge- and reasoning-oriented domains: Reasoning, STEM, and Humanities. Each domain contains 10 two-turn questions. We report turn-level domain scores and an overall average across all valid domain–question–turn scores. A score of -1 indicates an invalid or missing judgment and is excluded from all means. For each random-removal budget, we pool the valid scores from the three sampled masks before computing the mean.

3. Diagnostic Findings

We report one macro analysis and two micro analyses. The macro analysis looks at aggregate scores to ask how much

Table 1. **Speech MT-Bench scores (higher is better) for reference systems, budget-level means, and individual masks.** Mean rows pool three sampled masks; mask rows show individual samples. Percentages are relative to Full KAME. Reason. and Human. abbreviate Reasoning and Humanities; T1 and T2 denote the first and second turns of each two-turn dialogue.

Variant	Reason. T1	Reason. T2	STEM T1	STEM T2	Human. T1	Human. T2	Overall
Full KAME ($K = 0$)	4.90	5.00	6.30	5.40	6.70	6.20	5.75
Moshi (S2S)	1.05 (21%)	1.05 (21%)	2.45 (39%)	2.45 (45%)	1.65 (25%)	1.65 (27%)	1.72 (30%)
$K = 3$ mean	5.97 (**122%**)	5.43 (**109%**)	5.57 (**88%**)	6.63 (**123%**)	6.55 (**98%**)	6.18 (**100%**)	6.05 (**105%**)
$K = 5$ mean	4.57 (**93%**)	4.70 (**94%**)	5.22 (**83%**)	6.13 (**114%**)	5.13 (77%)	4.83 (78%)	5.10 (**89%**)
$K = 7$ mean	4.47 (**91%**)	5.07 (**101%**)	4.30 (68%)	4.73 (**88%**)	4.38 (65%)	3.86 (62%)	4.47 (78%)
$K = 10$ mean	3.33 (68%)	3.03 (61%)	2.60 (41%)	3.37 (62%)	2.87 (43%)	2.79 (45%)	3.00 (52%)
$K = 3$ mask A	5.30	4.50	5.20	6.40	6.05	5.17	5.44
$K = 3$ mask B	6.80	6.10	6.45	7.20	6.50	6.40	6.58
$K = 3$ mask C	5.80	5.70	5.05	6.30	7.10	6.94	6.14
$K = 5$ mask A	4.60	4.10	5.30	5.90	5.20	5.75	5.12
$K = 5$ mask B	4.40	4.80	5.95	7.30	5.10	4.75	5.38
$K = 5$ mask C	4.70	5.20	4.40	5.20	5.10	4.11	4.80
$K = 7$ mask A	5.80	5.00	4.50	5.10	4.40	4.00	4.80
$K = 7$ mask B	5.10	5.60	5.00	5.10	5.15	5.00	5.16
$K = 7$ mask C	2.50	4.60	3.40	4.00	3.60	2.56	3.46
$K = 10$ mask A	4.40	3.10	2.60	2.60	2.30	2.00	2.83
$K = 10$ mask B	2.90	3.40	3.00	4.80	3.70	3.20	3.50
$K = 10$ mask C	2.70	2.60	2.20	2.70	2.60	3.22	2.66

front-end depth can be removed. The micro analyses look at dialogue-level differences to ask what the failures look like and whether similar scores imply similar behavior.

3.1. Macro: Budgeted Removal Preserves Tandem Benefit

For the macro analysis, we look at aggregate scores to ask how much front-end depth can be removed before the tandem benefit starts to break.

Table 1 summarizes the budget-quality frontier. The main pattern is a recoverable frontier rather than immediate collapse. With $K = 3$ mean, the reduced front-end remains in the same score range as Full KAME (105% overall); we do not treat the small increase as a stable improvement, but it indicates a slack region that recovery fine-tuning can absorb. With $K = 5$ and $K = 7$ means, the reduced front-end retains 89% and 78% of the Full KAME overall score. At $K = 10$ mean, quality drops sharply to 52% overall, but the reduced tandem model still remains above the Moshi S2S baseline in every domain–turn column. The cliff is therefore relative to the full front-end, not a collapse of the tandem advantage.

Selective preservation across response types. Moderate removal preserves some response types more reliably than others. At $K = 7$ mean, Reasoning preserves 96% (91% on turn 1 and 101% on turn 2) of the Full KAME domain score, while STEM and Humanities preserve 77% (68% on turn 1 and 88% on turn 2) and 64% (65% on turn 1 and 62% on turn 2). Thus, the model has not collapsed, but response types already degrade at different rates. We do not attribute this to domain-specific layers; the narrower conclusion is that reasoning-oriented behavior remains more recoverable under moderate reduction, whereas longer structured responses appear more fragile. Section 3.2 examines this response-level pattern through the length distribution of low-score responses.

3.2. Micro 1: Low Scores Are Often Long-but-Wrong

For the first micro analysis, we look at low-score dialogue turns to see whether the model stops responding or keeps producing long but wrong answers.

Table 2 reports the low-score rate and the length distribution within low-score responses. We define a low-score response as a valid turn-level response with judge score at most 2. We use answer length only to characterize the failure shape: Silent-like responses are empty or very short (0–19 words), while Long responses contain at least 151 words.

The $K = 7$ setting shows the failure shape before the overall cliff. At this budget, aggregate quality is still partly preserved, yet low-score responses already differ by domain: among low-score cases, Long responses account for 43% in Reasoning, 65% in STEM, and 76% in Humanities. Thus, the macro pattern in Table 1 has a response-level counterpart: moderate removal is more recoverable for reasoning-style responses, whereas longer structured responses are more fragile.

At $K = 10$, the same failure shape becomes much stronger. The low-score rate rises to 61%, but the model rarely fails by becoming silent: only 15% of low-score responses are

Table 2. **Failure shape among low-score responses.** Low-score means judge score ≤ 2; Silent-like is 0–19 words and Long is $\geq$151 words. Rows with $K > 0$ pool three sampled masks; invalid -1 judgments are excluded.

Group	Low-score rate	Silent-like among low	Long among low
Full KAME	30%	11%	50%
$K = 7$ All	33%	12%	60%
Reason.	35%	10%	43%
STEM	33%	10%	65%
Human.	29%	18%	76%
$K = 10$ All	**61%**	15%	**66%**
Reason.	60%	14%	53%
STEM	57%	3%	**85%**
Human.	66%	26%	62%

Silent-like, whereas 66% are Long. STEM is the clearest case, with 85% of low-score responses falling into the Long category. This suggests that the cliff reflects less a loss of response generation than a loss of sustained correctness, coherence, and oracle-guided response construction.

3.3. Micro 2: Similar Scores Hide Item-Level Differences

For the second micro analysis, we ask whether similar aggregate scores imply similar retained dialogue behavior. We compare same-budget mask pairs whose domain means differ by at most 0.3 points, and then examine matched question–turn scores within that domain.

Even when domain-level scores are close, item-level behavior can diverge substantially. Across the non-cliff budgets ($K \leq 7$), about one third of matched question–turns differ by at least three judge points, with a maximum gap of eight points; including $K = 10$ gives the same qualitative pattern.

For example, two $K = 3$ masks have close STEM scores in Table 1: mask A scores 5.20/6.40 on STEM T1/T2, while mask C scores 5.05/6.30. Yet on one STEM follow-up item asking for edge cases, mask A receives 9 while mask C receives 2. Both responses are long, but the lower-scoring response becomes confused and errorful.

This is not evidence that particular layers encode particular question types. Rather, it is a cautionary diagnostic: under the same removal budget, similar aggregate or domain-level scores can still hide which dialogue behaviors remain reliable. Future reduction methods should therefore track micro-level behavior in addition to overall score.

4. Discussion and Limitations

The diagnostic points to a recoverable but bounded form of front-end capacity. Moderate removal should not be read as evidence that removed layers are simply useless. Rather, tandem delegation changes the workload: once a back-end supplies semantic guidance, recovery fine-tuning can absorb moderate perturbations to front-end depth. The fact that the reduced $K = 10$ model remains above the Moshi S2S baseline further suggests that the tandem benefit persists under substantial reduction.

One possible reason for the recoverable region is that the lower-level speech representation and stream interface already provide a scaffold for locally grounded behavior. In a Moshi-style front-end, Mimi (Défossez et al., 2024) converts audio into discrete codec tokens, giving the model a structured speech-token interface. This remains an interpretation rather than a causal claim, since we do not ablate Mimi, the codec interface, or the protected lower layers.

The capacity boundary reveals what remaining depth is needed for. The long-but-wrong failures suggest that the front-end is not merely a speech realization module: it can often continue responding, but loses the ability to maintain correctness, coherence, and alignment with evolving back-end guidance. The hard compression problem is therefore to preserve the online control needed to turn an oracle stream into a stable spoken trajectory.

Several limitations qualify the result. Budgeted random removal is a diagnostic protocol, not a final compression method, and three sampled masks per budget are not enough to treat small differences as stable improvements. The evaluation relies on speech MT-Bench with an LLM judge, which is scalable but does not replace human evaluation of naturalness, interruption handling, intelligibility, responsiveness, or preference. Finally, latency in full-duplex speech systems is not captured by a single end-of-utterance metric; real-time factor, memory footprint, and GPU utilization would complement the quality analysis.

Future work should move from measuring this sampled-removal frontier to pushing it outward with structured layer selection, recovery distillation, and oracle-aware compression objectives, while evaluating both aggregate scores and response-level spoken-dialogue behavior.

5. Conclusion

We studied layer removal as a diagnostic of front-end capacity in tandem speech dialogue. Moderate removal preserves much of the KAME-style tandem benefit after oracle-conditioned recovery fine-tuning, while aggressive removal exposes a bounded capacity frontier. Failures are often long-but-wrong rather than silent: reduced front-ends continue speaking but lose correctness, coherence, and oracle-guided control. Similar aggregate scores can also hide different item-level dialogue behaviors. These findings set a useful direction for future structured reduction: smaller real-time front-ends that preserve online response control.

References

Chien, C.-M., Orsini, M., Kharitonov, E., Zeghidour, N., Livescu, K., and Défossez, A. MoshiRAG: Asynchronous knowledge retrieval for full-duplex speech language models. *International Conference on Machine Learning*, 2026.

Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Dorszewski, T., Jacobsen, A. K., Tětková, L., and Hansen, L. K. How redundant is the transformer stack in speech representation models? In *ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Fan, A., Grave, E., and Joulin, A. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

Kim, B.-K., Kim, G., Kim, T.-H., Castells, T., Choi, S., Shin, J., and Song, H.-K. Shortened LLaMA: Depth pruning for large language models with comparison of retraining methods. *arXiv preprint arXiv:2402.02834*, 2024.

Kuroki, S., Kubo, Y., Akiba, T., and Tang, Y. KAME: Tandem architecture for enhancing knowledge in real-time speech-to-speech conversational ai. In *ICASSP 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19362–19366. IEEE, 2026.

Men, X., Xu, M., Zhang, Q., Yuan, Q., Wang, B., Lin, H., Lu, Y., Han, X., and Chen, W. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, 2025.

Wang, X., Li, Y., Fu, C., Zhang, Y., Shen, Y., Xie, L., Li, K., Sun, X., and Ma, L. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *International Conference on Machine Learning*, 2025.

Xie, Z. and Wu, C. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

Zhong, L., Wan, F., Chen, R., Quan, X., and Li, L. BlockPruner: Fine-grained pruning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5065–5080, 2025.