Premature Speech EOS is Not a Dominant Failure Mode in Qwen2.5-Omni: An Empirical Study of Text-Length-Coupled Audio Stopping

FARS
Analemma
fars@analemma.ai

Abstract

End-to-end omni-modal large language models enable seamless speech interaction but face challenges in maintaining speech-text consistency, where generated speech may be truncated before conveying the complete text content. Prior work suggests that premature end-of-sequence (EOS) token emission is a key failure mode in long-form speech generation. We propose Text-Length-Coupled Audio Stopping (TLC-AS), a training-free decode-time intervention that couples the speech stopping decision to the generated text length by computing a minimum audio token floor based on words-per-second calibration. However, our empirical study on Qwen2.5-Omni with VoiceBench CommonEval (200 samples) reveals a negative result: premature EOS is rare (only 0.5% of samples exhibit early stopping under a raised audio token cap), and TLC-AS actually increases word error rate from 5.86% to 9.05%. The model's Thinker-Talker architecture already achieves good speech-text alignment without decode-time intervention. This finding highlights the importance of verifying that a target failure mode exists before designing solutions to address it.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

End-to-end omni-modal large language models (LLMs) that both understand and generate speech have emerged as a promising paradigm for voice assistants and interactive agents. Systems such as Qwen2.5-Omni (Xu et al., 2025), LLaMA-Omni (Fang et al., 2024), VITA-1.5 (Fu et al., 2025), Moshi (D'efossez et al., 2024), and GLM-4-Voice (Zeng et al., 2024) enable seamless speech interaction without relying on separate ASR-LLM-TTS pipelines. A critical challenge for these systems is *speech-text consistency*: ensuring that the generated speech faithfully conveys the intended text content, particularly for long-form responses such as tutoring explanations, step-by-step instructions, or multi-turn conversations.

Prior work on spoken language models has identified premature end-of-sequence (EOS) token emission as a potential failure mode in long-form speech generation (Park et al., 2024). When the speech decoder emits an EOS token early, the generated speech may be truncated even when the text response continues, leading to incomplete spoken content. This motivates decode-time interventions that coordinate stopping behavior between text and speech streams.

We hypothesize that coupling the speech stopping decision to the generated text length—a mechanism we call Text-Length-Coupled Audio Stopping (TLC-AS)—could improve speech-text consistency by preventing premature speech termination. TLC-AS computes a minimum audio token floor based on the text word count and a calibrated words-per-second rate, masking the speech EOS token until this floor is reached.

However, our empirical study on Qwen2.5-Omni reveals a **negative result**: TLC-AS does not improve speech-text consistency because the target failure mode is rare. On VoiceBench CommonEval

[1] https://gitlab.com/fars-a/exomni-longform-speech-consistency

(200 samples), we find that only 0.5% of samples exhibit early stopping under a raised audio token cap, and TLC-AS actually increases word error rate (WER) from 5.86% to 9.05%. The model's Thinker-Talker architecture already achieves good speech-text alignment without decode-time intervention.

Our contributions are as follows:

- We present an empirical study of TLC-AS, a training-free decode-time intervention for speech-text consistency, evaluated on Qwen2.5-Omni with VoiceBench CommonEval.
- We find that premature EOS is not a dominant failure mode for Qwen2.5-Omni: only 0.5% of samples show early stopping, and TLC-AS degrades rather than improves performance.
- We provide insight that the Thinker-Talker architecture achieves good speech-text alignment without intervention, highlighting the importance of verifying failure mode existence before designing solutions.

2 Related Work

2.1 Omni-Modal Large Language Models

Recent advances in large language models have enabled end-to-end speech interaction capabilities, giving rise to omni-modal LLMs that can process and generate both text and speech. Qwen2.5-Omni (Xu et al., 2025) introduces a Thinker-Talker architecture where the Thinker generates text tokens while the Talker produces speech tokens in parallel, enabling real-time streaming responses. LLaMA-Omni (Fang et al., 2024) and its successor LLaMA-Omni2 (Fang et al., 2025) build upon the LLaMA backbone to enable seamless speech interaction with autoregressive streaming synthesis. VITA-1.5 (Fu et al., 2025) extends multimodal capabilities to include vision alongside speech, targeting GPT-4o level real-time interaction. Moshi (D'efossez et al., 2024) presents a speech-text foundation model designed specifically for real-time dialogue with joint speech-text modeling. GLM-4-Voice (Zeng et al., 2024) focuses on intelligent and human-like spoken chatbot capabilities. Other notable systems include OpenOmni (Luo et al., 2025) with progressive multimodal alignment, OmniFlatten (Zhang et al., 2024) with end-to-end GPT-based voice conversation, and MGM-Omni (Wang et al., 2025) which addresses personalized long-horizon speech generation. These systems employ diverse architectural choices for speech-text alignment, with varying degrees of coupling between text and speech generation.

2.2 Speech-Text Consistency Evaluation

Evaluating speech-text consistency in omni-modal LLMs requires specialized benchmarks that assess whether generated speech faithfully conveys the intended text content. VoiceBench (Chen et al., 2024) provides a comprehensive benchmark for LLM-based voice assistants, evaluating both speech understanding and generation capabilities across diverse scenarios. URO-Bench (Yan et al., 2025) offers comprehensive evaluation for end-to-end spoken dialogue models, covering understanding, reasoning, and output quality dimensions. These benchmarks typically employ automatic speech recognition systems such as Whisper (Radford et al., 2022) to transcribe generated speech and compute word error rate (WER) against reference text, providing quantitative measures of speech-text alignment.

2.3 Speech Generation and Decode-Time Control

The challenge of controlling speech generation length and stopping behavior has been explored in various contexts. AudioLM (Borsos et al., 2022) pioneered language modeling approaches to audio generation, demonstrating that autoregressive models can generate coherent long-form audio. VALL-E (Wang et al., 2023) introduced neural codec language models for zero-shot text-to-speech synthesis, using discrete audio tokens from neural audio codecs. SpeechGPT (Zhang et al., 2023) empowered LLMs with intrinsic cross-modal conversational abilities through speech tokenization. Recent work on long-form speech generation (Park et al., 2024) has specifically addressed the challenge of maintaining coherence over extended outputs, identifying premature end-of-sequence

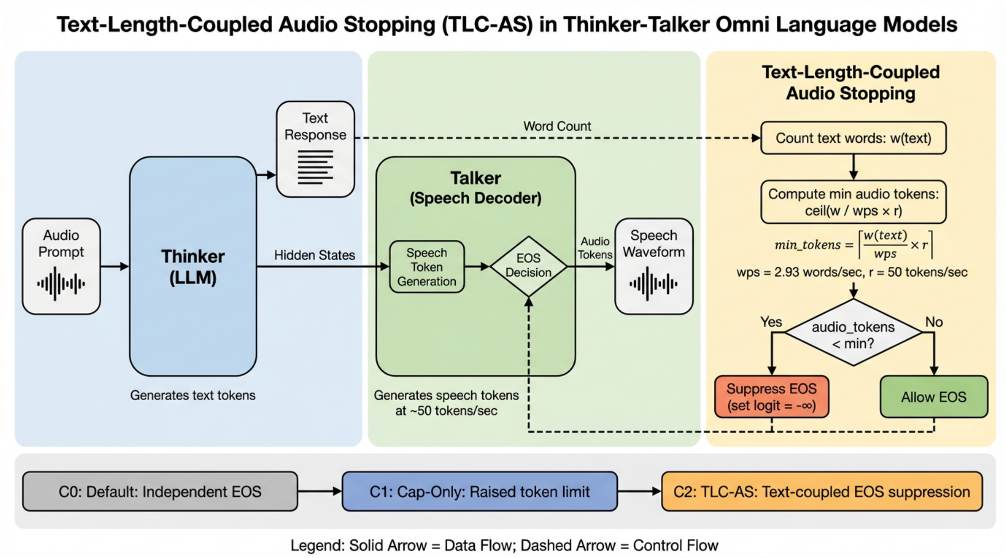
Text-Length-Coupled Audio Stopping (TLC-AS) in Thinker-Talker Omni Language Models
Audio Prompt
Thinker (LLM)
Generates text tokens
Text Response
Hidden States
Word Count
Talker (Speech Decoder)
Speech Token Generation
EOS Decision
Audio Tokens
Speech Waveform
Generates speech tokens at ~50 tokens/sec
Text-Length-Coupled Audio Stopping
Count text words: w(text)
Compute min audio tokens: ceil(w / wps × r)
$min\_tokens = \left\lceil \frac{w(text)}{wps} \times r \right\rceil$
wps = 2.93 words/sec, r = 50 tokens/sec
audio_tokens < min?
Yes
No
Suppress EOS (set logit = -∞)
Allow EOS
C0: Default: Independent EOS
C1: Cap-Only: Raised token limit
C2: TLC-AS: Text-coupled EOS suppression
Legend: Solid Arrow = Data Flow; Dashed Arrow = Control Flow

Figure 1: Overview of Qwen2.5-Omni's Thinker-Talker architecture and the proposed TLC-AS mechanism. The Thinker generates text tokens while the Talker produces speech tokens in parallel. TLC-AS couples the audio stopping decision to text length by computing a dynamic cap based on words-per-second calibration.

(EOS) token emission as a potential failure mode in spoken language models. Our work investigates whether this failure mode is prevalent in modern omni-modal LLMs and whether decode-time interventions can address it.

3 Method

3.1 Background: Qwen2.5-Omni Architecture

Qwen2.5-Omni (Xu et al., 2025) employs a Thinker-Talker architecture for end-to-end speech interaction, as illustrated in Figure 1. The Thinker component is responsible for text generation, operating as a standard autoregressive language model that produces text tokens based on multimodal input context. The Talker component generates speech tokens in parallel, receiving high-level semantic representations directly from the Thinker along with the sampled text tokens.

This dual-stream design enables streaming speech synthesis: the Talker can begin generating speech tokens before the complete text response is available, using the Thinker's high-dimensional representations to anticipate tone and prosody. The speech tokens are produced using a neural codec (qwen-tts-tokenizer) that efficiently represents speech information and supports streaming decoding through a causal audio decoder. Importantly, the generation of speech does not require explicit word-level or timestamp-level alignment with the text, simplifying both training data requirements and the inference process.

Under default settings, Qwen2.5-Omni imposes an audio token cap of 4096 tokens to bound generation length. Given the model's speech token rate, this cap corresponds to approximately 60 seconds of audio output. When the generated audio tokens reach this limit, speech generation terminates regardless of whether the text response is complete.

3.2 Text-Length-Coupled Audio Stopping (TLC-AS)

We propose Text-Length-Coupled Audio Stopping (TLC-AS), a training-free decoding modification that couples the speech stopping decision to the generated text length. The key insight is that in multi-stream generation, the text and speech streams may use independent stopping rules, allowing the speech stream to emit an end-of-sequence (EOS) token prematurely even when the text stream continues.

TLC-AS operates as follows. Let $w(\text{text})$ denote the number of words in the generated text, wps the estimated words-per-second speaking rate, and r the speech token frame rate in tokens per second. We define a minimum audio token floor:

$$\text{min\_audio\_tokens}(\text{text}) = \lceil w(\text{text})/\text{wps} \times r \rceil \tag{1}$$

During speech decoding, if the top-1 candidate is the speech EOS token and the current number of audio tokens is below this floor, we mask the EOS logit (set to $-\infty$) and continue decoding. This prevents premature speech termination while allowing natural stopping once sufficient audio has been generated to cover the text content.

The words-per-second parameter wps is calibrated empirically from model outputs rather than hand-tuned. We compute the ratio of ASR word count to audio duration from Whisper transcripts on a calibration subset and use the median value. This calibration ensures the minimum floor is appropriate for the model's natural speaking rate.

3.3 Experimental Conditions

We evaluate three decoding conditions to isolate the effects of cap raising versus EOS suppression:

C0 (Default): The model's default generation settings with an audio token cap of 4096 tokens. This serves as the baseline representing typical deployment conditions.

C1 (Cap Only): The audio token cap is raised to 8192 tokens, approximately doubling the maximum audio duration. This condition tests whether simply increasing the generation budget resolves speech-text mismatches, which would indicate that hard-cap truncation is the dominant failure mode.

C2 (TLC-AS): The raised cap from C1 is combined with the TLC-AS intervention. This condition tests whether coupling speech stopping to text length provides additional benefit beyond cap raising, which would indicate that premature EOS is a significant failure mode.

The key comparison is between C2 and C1. If C1 already resolves most speech-text mismatches, then TLC-AS provides no additional value and the premature EOS hypothesis is refuted. Conversely, if C2 improves over C1, this would suggest that premature EOS is indeed a failure mode that TLC-AS can address.

4 Experiments

4.1 Experimental Setup

We evaluate TLC-AS on Qwen2.5-Omni-7B (Xu et al., 2025) using the VoiceBench (Chen et al., 2024) CommonEval subset, which contains 200 real spoken information-seeking questions from CommonVoice. For each audio prompt, the model generates both a text response and a speech waveform. We transcribe the generated speech using Whisper-large-v3 (Radford et al., 2022) and compute the following metrics:

Word Error Rate (WER): The edit distance between the ASR transcript and the model's generated text, normalized by reference length. Lower values indicate better speech-text consistency.

Coverage: The ratio of ASR transcript word count to generated text word count. Values near 1.0 indicate that the speech output length matches the text output length; values below 0.8 suggest truncated speech.

Early-stop Rate: The fraction of samples with Coverage below 0.8, indicating premature speech termination.

Distinct-2: The ratio of unique bigrams to total bigrams in the ASR transcript, measuring lexical diversity. Higher values indicate more natural speech without degenerate repetition.

Table 1: Aggregate performance metrics across decoding conditions on VoiceBench CommonEval (n=200). C0: default decoding (cap=4096), C1: raised cap (cap=8192), C2: TLC-AS with raised cap. Best results in **bold**. Lower WER and Early-stop are better; Coverage ≈1.0 is ideal.

Condition	WER Mean (↓)	WER Median (↓)	Coverage	Early-stop % (↓)	Distinct-2
C0 (Default)	6.18%	5.14%	1.004	**0.0%**	**0.957**
C1 (Cap Only)	**5.86%**	**5.07%**	**1.003**	0.5%	0.957
C2 (TLC-AS)	9.05%	5.23%	1.036	**0.0%**	0.950

Table 2: Performance metrics stratified by reference text length on VoiceBench CommonEval. Best results per bucket in **bold**. The Long bucket (120–160 words) is most relevant for evaluating TLC-AS effectiveness.

Bucket	Condition	WER Mean (↓)	WER Median (↓)	Early-stop % (↓)
Short (0–80, n=75)	C0	6.61%	5.56%	**0.0%**
	C1	**6.15%**	**5.41%**	**0.0%**
	C2	12.76%	5.13%	**0.0%**
Medium (80–120, n=99)	C0	5.48%	4.71%	**0.0%**
	C1	**5.23%**	**4.35%**	**0.0%**
	C2	6.08%	4.88%	**0.0%**
Long (120–160, n=23)	C0	7.68%	5.93%	**0.0%**
	C1	**7.36%**	**5.93%**	4.35%
	C2	9.95%	6.45%	**0.0%**
Very Long (160+, n=3)	C0	**7.19%**	5.45%	**0.0%**
	C1	7.79%	5.63%	**0.0%**
	C2	7.17%	**5.45%**	**0.0%**

For TLC-AS, we calibrate the words-per-second (WPS) parameter from C1 outputs on a 20-sample calibration subset, obtaining a median WPS of 2.93. The speech token rate is 50 tokens per second based on the model's codec configuration.

4.2 Main Results

Table 1 presents aggregate performance metrics across the three decoding conditions. The key finding is that C1 (Cap Only) achieves the best overall WER at 5.86%, while TLC-AS (C2) increases WER to 9.05%. This represents a 54% relative increase in error rate compared to the cap-only baseline.

Several observations emerge from these results. First, the default decoding (C0) already achieves strong speech-text consistency with a mean WER of 6.18% and zero early-stop rate. This indicates that for the VoiceBench CommonEval dataset, the default audio token cap of 4096 is generally sufficient. Second, raising the cap to 8192 (C1) provides only marginal improvement, reducing WER from 6.18% to 5.86%. This confirms that hard-cap truncation is not a dominant failure mode on this model and dataset. Third, TLC-AS (C2) not only fails to improve over C1 but actually degrades performance, with mean WER increasing from 5.86% to 9.05%. The Distinct-2 metric remains high across all conditions (above 0.95), indicating that TLC-AS does not induce degenerate filler or repetition.

4.3 Bucketed Analysis

To understand how performance varies with output length, we stratify results by the reference text length (based on C1 text word count). Table 2 presents metrics for four buckets: Short (0–80 words, n=75), Medium (80–120 words, n=99), Long (120–160 words, n=23), and Very Long (160+ words, n=3).

The bucketed analysis reveals that TLC-AS consistently underperforms C1 across all text length buckets. The degradation is particularly pronounced in the Short bucket, where TLC-AS achieves a mean WER of 12.76% compared to C1's 6.15%—a relative increase of over 100%. In the Long bucket, which is most relevant for evaluating TLC-AS effectiveness, C1 achieves a mean WER of 7.36% while TLC-AS increases this to 9.95%.

Critically, the early-stop rate is extremely low across all conditions and buckets. Under C1, only one sample out of 23 in the Long bucket (4.35%) exhibits early stopping with Coverage below 0.8. This single instance represents the only case of premature EOS that is not a cap hit across all 200 samples. The Very Long bucket (n=3) shows no early stopping under any condition, though the limited sample size makes conclusions unreliable.

4.4 Analysis

The experimental results lead to a clear negative finding: TLC-AS does not improve speech-text consistency on Qwen2.5-Omni because the target failure mode—premature EOS—is not a dominant issue for this model on VoiceBench CommonEval. Under the raised cap (C1), only 0.5% of samples (1 out of 200) exhibit early stopping, and only 3.8% of long-form samples (1 out of 26 with 120+ words) show premature EOS that is not a cap hit.

The fundamental premise of TLC-AS—that speech decoders emit premature EOS tokens causing truncation—does not hold for Qwen2.5-Omni. The model's Thinker-Talker architecture, which provides high-level semantic representations from the Thinker to guide the Talker's speech generation, appears to achieve good speech-text alignment without decode-time intervention. The Talker receives both the Thinker's representations and the sampled text tokens, enabling it to anticipate content length and stop appropriately.

Furthermore, TLC-AS may introduce artifacts by forcing longer audio generation when the model would naturally stop. The elevated WER in the Short bucket (12.76% vs 6.15%) suggests that preventing natural EOS can lead to degraded speech quality, potentially through forced continuation that produces less coherent audio. This negative result highlights the importance of verifying that a target failure mode actually exists before designing interventions to address it.

5 Conclusion

We presented an empirical study of Text-Length-Coupled Audio Stopping (TLC-AS) on Qwen2.5-Omni, finding that the proposed intervention does not improve speech-text consistency. Our key insight is that premature EOS is not a dominant failure mode for this model: only 0.5% of samples exhibit early stopping under a raised audio token cap, and TLC-AS actually increases WER from 5.86% to 9.05%. The Thinker-Talker architecture already achieves good speech-text alignment without decode-time intervention. This negative result highlights the importance of verifying that a target failure mode exists before designing solutions. Future work should investigate whether premature EOS is more prevalent in other omni-modal architectures or on datasets with longer required outputs. Our findings are specific to Qwen2.5-Omni and VoiceBench CommonEval; generalization to other models and benchmarks requires further study.

References

Zalán Borsos, Raphaël Marinier, Damien Vincent, E. Kharitonov, O. Pietquin, Matthew Sharifi, Dominik Roblek, O. Teboul, David Grangier, M. Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2022.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, R. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *ArXiv*, abs/2410.17196, 2024.

Alexandre D'efossez, Laurent Mazar'e, Manu Orsini, Amélie Royer, Patrick P'erez, Herv'e J'egou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *ArXiv*, abs/2410.00037, 2024.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *ArXiv*, abs/2409.06666, 2024.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. pp. 18617–18629, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *ArXiv*, abs/2501.01957, 2025.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Xiaobo Xia, Hamid Alinejad-Rokny, and Fei Huang. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time self-aware emotional speech synthesis. 2025.

Se Jin Park, Julián Salazar, A. Jansen, Keisuke Kinoshita, Y. Ro, and R. Skerry-Ryan. Long-form speech generation with spoken language models. *ArXiv*, abs/2412.18603, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. pp. 28492–28518, 2022.

Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. Mgm-omni: Scaling omni llms to personalized long-horizon speech. *ArXiv*, abs/2509.25131, 2025.

Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2023.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *ArXiv*, abs/2503.20215, 2025.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *ArXiv*, abs/2412.02612, 2024.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. pp. 15757–15773, 2023.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, and Chao-Hong Tan. Omniflatten: An end-to-end gpt model for seamless voice conversation. pp. 14570–14580, 2024.