Evaluating ASR Quality at Scale on TV Entertainment Platforms

Adeep Hande Kishorekumar Sundararajan, Yidnekachew Endale,
Akshatha Bapu KrishnaSwamy, Sachin Dabral, Dawn Reed, Michael Pereira
Applied AI Research, Comcast
Washington D.C, USA
adeep_hande2@comcast.com
kishore_kumar@comcast.com

Abstract

Evaluating automatic speech recognition (ASR) quality in production is a fundamentally challenging problem, owing to the volume and variability of real-time user interactions, thus making traditional methods impractical, and ground-truth annotations are typically unavailable at scale. We propose a large-scale evaluation pipeline that enables ASR quality evaluation via weakly supervised estimations derived from multiple independent ASR engines. The proxy estimations' outputs are used to compute consensus-driven metrics such as word error rate (WER), character error rate (CER), and semantic divergence – to estimate transcription quality across varied query categories. These signals form the basis for a large language model (LLM)-driven insight agent, and custom tools to detect failure patterns, extract usage trends and behaviors, and provide actionable insights. This framework serves as a final layer of evaluation, which acts as an independent module compatible with any ASR system, regardless of its underlying architecture. By processing millions of queries periodically without manual intervention, this solution provides actionable insights that support the research and product team with valuable feedback for continuous improvement.

1 Introduction

Modern large scale voice interaction platforms, handle tens of millions of user voice commands/utterances across a wide range of domains (sports, reality TV, movies, etc.). As the deployment scale continues to grow, evaluating ASR performances on real-time traffic becomes increasingly difficult. Ensuring a high ASR accuracy at scale - and quickly fixing any model shortcomings - is crucial for a good user experience. Conventional evaluation metrics such as Word Error Rate (WER) and Character Error Rate (CER) rely on the presence of ground-truths transcriptions. In production scale, however, obtaining manual reference transcripts for every user utterance is not feasible - it is time-consuming, resource-intensive, and quite impractical to do at scale (Yuksel et al.). Additionally, being completely reliant on intermittent manual edits or curated test sets tend to be insufficient for continual monitoring of utterances. These drawbacks push the need for a referenceless evaluation approach on raw ASR outputs, without human transcripts.

To bridge this gap, we propose a consensus-based weak supervision framework for ASR evaluation. analyzing a significant fraction of real user queries daily. Rather than relying on single ground-truth transcript, our system leverages agreements across multiple audio signals to gauge transcript quality, and fuse outputs from multiple ASR engines to produce a pseudo-transcription based on high-confidence matches and its supplementary metrics (Prakash et al., 2025). If the majority of annotators align with the same transcript on an utterance, it is likely correct; conversely, if there is no clear majority, the framework relies on semantic divergence via clustering to assess the estimated transcript. This ensemble strategy can largely help mitigate ASR errors and improve confidence on the system by not being completely reliant on any single ASR engine for ground truth (Waheed et al., 2025). Such weak supervision approach yields a reference quality estimate, which can be further used to derive other traditional metrics such as WER, and CER for interpretation. This allows us to detect any potential issues on live traffic automatically. For instance, if any high-traffic utterance has a lower match rate with any of the systems consistently, it can be further analyzed. Crucially, this proxy estimation approach allows us to scale this evaluation to the production traffic and requires no manual annotations.

Beyond just flagging match rates and error rates, our framework emphasizes interpretation and aims

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 378–383

to provide insight into the ASR performance. We integrate a large language model (LLM) into the evaluation pipeline to turn the ASR transcription data and its associated metrics into meaningful trend analysis (Yuan et al., 2024; Li et al., 2024). The LLM ingests large batches of different types of utterances and their scores to identify patterns and explainable causes of errors. For instance, it can automatically surface errors involving recent movie titles (Ex, *"Minions: The Rise of Gru"*) or that recognition drops for certain specific instances, while also recognizing patterns and trends over a period. This LLM-driven insight agent enables the product and research teams to quickly interpret the more prevalent issues (specific vocabulary gaps, content ambiguity) and monitor how the trends evolve over a specific period of time, in response to ASR model retraining and updates. The combination of weakly supervised quantitative signals and insights makes the framework highly practical and interpretable for efficient decision making.

The system analyzes a significant portion of the voice traffic daily to ensure efficient ASR quality and catch new issues early. The system gets fed a predefined type of utterance classes (*Long-tail*, *Long-Audio*, *Short-Audio*, etc), and eventually gives out insights to the teams after deriving the transcript estimations. This system effectively serves as additional feedback for model retraining and evaluation. This hybrid approach, combining qualitative and quantitative estimations for utterances, establishes a scalable evaluation framework for quality assurance teams. We demonstrate that this approach yields a high-confidence assessment on production voice utterances, estimating the true error rates while mitigating the need for manual transcription. Our approach bridges a crucial gap in industry-scale speech recognition: ensuring interpretable insights even without any ground-truths, enhancing the voice experience for millions of our customers.

Our contributions are as follows: (1) a consensus-based framework for generating pseudo-references from multiple independent ASR engines, enabling referenceless quality estimation at production scale, (2) an LLM-driven insight extraction pipeline that surfaces interpretable failure patterns and trends without manual annotation, and (3) a demonstration of this framework on a large-scale TV entertainment voice platform, processing millions of utterances periodically across multiple query categories.

2 Related Work

Referenceless ASR Evaluation: Conventional ASR evaluation is dependent on the availability of ground-truth metrics to evaluate metrics such as WER and CER, which limits their applicability in production environments. Researchers proposed NoRefer, a referenceless quality metric that leverages semi-supervised language model fine-tuning with contrastive learning to estimate transcription quality without reference transcripts (Yuksel et al.). Similarly, other approaches to explore robust approximation methods for ASR metrics show that proxy-based estimations can closely approximate the true error rate.(Waheed et al., 2025). Our work differs from these approaches in that we do not train a dedicated quality estimation model, but instead rely on consensus agreement across multiple independent ASR engines to construct a proxy reference, which eventually derives the metrics directly.

Multi-System Consensus and LLM-based evaluation: (Fiscus, 1997) introduced ROVER, aligning and voting across multiple ASR hypotheses to reduce the error rates, and (Prakash et al., 2025) proposed a multi-ASR fusion approach with speech-based language models for better pseudo-label generation. Outside these approaches, our work leverages the consensus signal specifically for evaluation — to assess transcription quality rather than to improve the transcription itself. LLMs have been effectively used as judges and evaluators in prior work (Fabbri et al., 2025; Zheng et al., 2023). We employ an LLM not to judge individual outputs but to analyze aggregate patterns across large volumes of ASR transcriptions, surfacing interpretable insights such as entity-level failures and category-specific trends (Yuan et al., 2024; Li et al., 2024).

3 Methodology

Our evaluation pipelines comprises of four stages: (1) audio sampling (data ingestion), (2) multi-ASR transcription and pseudo-reference generation, (3) WER/CER metric aggregation using weak labels, and (4) LLM-base insight generation, as shown in Fig 1.

3.1 Sampling and Categorization

A fixed daily percentage of real-time voice commands from our internal production traffic is sampled using uniform and stratified selection strategies. All audio samples are normalized to a wave-

Category	Insight Summary	Example + Actionable Step
Long-Audio Queries	Recognition quality degrades for utterances exceeding X seconds	"Play last night's episode of The SNL on Peacock…" → truncated *Apply segmentation/silence detection*
Music Queries	Trending Universal Music artist names are frequently misrecognized	"Play Billie Eilish" → "play belly eyelash" *Enhance named entity modeling*
TV Show Titles	Multi-word titles partially matched or reordered	"America's Got Talent: Fantasy League" → "American talent fantasy" *Add show title variants and alias support*
Short-Audio Queries	Ambiguous one-word queries misclassified	"Chicago" → routed to sports instead of TV show

Table 1: Representative LLM-generated ASR insight categories. All example utterances are synthetic and constructed to reflect observed patterns; no actual customer content is shown.

form with a specific frequency (16kHz). The utterances are classified into predefined categories for targeted evaluations, including, but not limited to:

- **Short-tail**: High-frequency, common user utterances
- **Long-tail**: Rare or ambiguous queries
- **Long-Audio/ Short-Audio**: Based on waveform duration exceeding a pre-defined threshold
- **Random**: Uniform sampling for control analysis

These categories ensure coverage across different types of domains and systems.

3.2 Multi-ASR Transcriptions and Pseudo-Label Consensus

Each sampled audio is normalized and independently transcribed by five ASR engines (Google Speech2Text API (Google Cloud, 2025), Amazon (Amazon Web Services, 2025), OpenAI Whisper (Radford et al., 2023), and an in-house model). We chose a majority threshold of three to balance confidence with coverage. For an utterance u_i, the system produces a transcription set, which are treated as weak annotators:

$$T = \{t_i^1, t_i^2, t_i^3, t_i^4, t_i^5\} \quad (1)$$

We incorporate a lightweight consensus mechanism to construct a proxy reference. If three or more systems produce near identical outputs, we adopt a ROVER-style approach (Fiscus, 1997). If there is no clear majority, we hypothesize that the most semantically similar transcription is the pseudo-transcription by computing a pairwise similarity score. Both of the cases would serve as weak labels for that utterance. Formally, for a given utterance u_i, we define the consensus-transcription, $\hat{T}$:

$$\hat{T} = \begin{cases} \text{Majority}(T), & \text{if } \exists\, t_i \in T \text{ such that} \\ & \sum_{j \neq i} \mathbb{1}(t_i = t_j) \geq 2 \\ \arg\max_{t_i \in T} \sum_{j \neq i} \cos\left(\mathbf{e}_i, \mathbf{e}_j\right), & \text{otherwise} \end{cases}$$

Where e_i is Sentence-BERT (Reimers and Gurevych, 2019) used to compute sentence-level embeddings for the latter. This ensemble approach avoids reliance on any single ASR model's output, which is valuable in production setting where no ground truth is available. Given each ASR transcription t_i^j for the utterance u_i and consensus $\hat{T}$, we compute proxy word error rate (WER) and character error rate (CER) using Levenshtein distance. These metrics are then aggregated across utterance-level classes (e.g, long-tail, long-audio, short-audio, etc.) and further serve their purpose for insight extraction via LLMs.

3.3 Insight Extraction

We leverage an LLM-based extraction pipeline to analyze weakly supervised ASR transcription data for detailed interpretation. The system currently leverages GPT-4 (Achiam et al., 2023) as

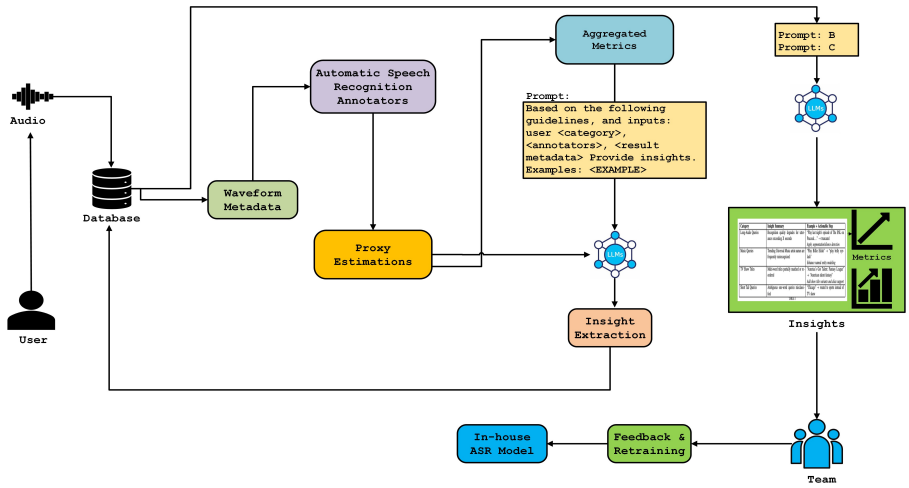

Figure 1: Overview of our end-to-end large scale audio evaluation framework.

the language model backend. The insights can be broadly categorized into three prompt-driven analysis stages, each specifying on a different granularity of evaluation. This multi-dimensional evaluation strategy is based on a task decomposition approach, enabling the LLM to handle complex tasks, improving clarity and performance by handling one task at a time (Khot et al.). In each stage, we provide structured guidelines and pre-formatted inputs to the LLM to ensure consistent and factual reporting. We describe the prompt design and functionality for each stage below, along with how we detect specific error patterns (such as specific failures and misrecognitions) to provide granular insights, examples of which are seen in Table 1.

- **Global Insight Summarization** (Prompt A): Produces an overall diagnostic summary, identifying dominant error-prone ASR engines, systemic biases or model drift.

- **Category-specific Analysis** (Prompt B): The system focuses on segmented analysis across several audio categories. We feed raw transcription data by category to the LLM along with several guidelines for evaluation. However, owing to the token length constraints, we chunk our data into several batches. To ensure consistency and context retention, we provide LLM outputs of prompt A to prompt B - a ReAct style chain-of-thought reasoning strategy that combines reasoning and tool usage in a loop (Yao et al., 2023).

- **Temporal Analysis** (Prompt C): The system leverages LLM outputs of prompt B by comparing daily batches of results, flagging regressions or improvements. This temporal framework enables easier interpretation and early detection of any drifts that were not captured during training.

A large language model (currently GPT-4) with structured input formats (JSON tables, labeled metrics, etc.). The prompts follow a few-shot templates to demonstrate expected outputs, ensuring reproducibility and interpretability. This staged prompting yields interpretable error analysis (see Table 1 for examples), which are otherwise hard to obtain from raw error rates.

4 Results and Evaluation

We evaluate the system across two axes: (1) the effectiveness of referenceless quality estimation via consensus-based metrics, and (2) the interpretability and utility of LLM-based insights for model diagnostics. For transcription quality estimation, the pipeline generates over 80% of its high-confidence sampled transcriptions through consensus transcriptions, and the rest by semantic embedding-based clustering when no majority exists, ensuring full coverage without relying on manual annotations. The LLM generation further generates interpretable and actionable insights, such as entity-level failures, numerical issues, etc, allowing QA to directly intervene to fix any potential issue prior to user reporting.

This framework serves as a continuous feedback loop for both product and research teams, as they offer interpretable diagnostics tied to specific instances of failures. Previously, evaluation

had to rely on limited test sets and sporadic manual checks. Now, our system continually provides high-quality data which can be further used for model fine-tuning and retraining. Furthermore, targeted entity-level and category-level performance metrics offer deeper insights for targeted model improvements. By combining human-readable insight with scalable metrics, the system enables both high-level interpretation and low-level correct action without the need for manual labeling, allowing for it to scale up to production traffic.

5 Conclusion

We present a referenceless large-scale ASR evaluation system, combining multi-ASR consensus, semantic clustering, and LLM-guided insight extraction. The system is currently scaled to a significant fraction of real-time utterances, requiring no manual annotations while generating valuable interpretable insights, metrics, surfacing model failures across audio categories. It serves as a critical feedback loop - enabling both research and product teams for model retraining and domain adaptations. By integrating the system into the evaluation pipeline, we provide a scalable human-centric interpretation of raw-ASR data, accelerating the evaluation cycle for model improvements, and ensuring reliable user experience in large-scale voice platforms.

6 Limitations

We acknowledge that due to the sensitive nature of production data, we do not disclose exact metrics and rather focus on qualitative findings and relative trends throughout this work. The framework has been demonstrated on English utterances only, and its effectiveness on other languages has not been explored. The LLM-based insight extraction currently relies on GPT-4, and its generalizability to other language models has not been evaluated.

7 Ethics Statement

All utterances in Table 1 and Fig 1 are synthetic and are constructed to reflect observed patterns. No actual customer audios, transcriptions, or any identifiable information is disclosed. The evaluation framework operates on anonymized production data, and no individual user behavior can be backtracked or identified from the reported findings. The use of commercial ASR engines and GPT-4 in this work is subject to the respective providers' terms of service and data handling policies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amazon Web Services. 2025. Amazon transcribe documentation. Accessed: 2025-06-23.

Francesco Fabbri, Gustavo Penha, Edoardo D'Amico, Alice Wang, Marco De Nadai, Jackie Doremus, Paul Gigioli, Andreas Damianou, Oskar Stål, and Mounia Lalmas. 2025. Evaluating podcast recommendations with profile-aware llm-as-a-judge. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, page 1181–1186, New York, NY, USA. Association for Computing Machinery.

Jonathan G Fiscus. 1997. 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*, pages 347–354. IEEE.

Google Cloud. 2025. Cloud speech-to-text documentation. Accessed: 2025-06-23.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Jiayang Li, Jiale Li, and Yunsheng Su. 2024. A map of exploring human interaction patterns with llm: Insights into collaboration and creativity. In *International Conference on Human-Computer Interaction*, pages 60–85. Springer.

Jeena Prakash, Blessingh Kumar, Kadri Hacioglu, Bidisha Sharma, Sindhuja Gopalan, Malolan Chetlur, Shankar Venkatesan, and Andreas Stolcke. 2025. Better pseudo-labeling with multi-asr fusion and error correction by speechllm. *Preprint*, arXiv:2506.11089.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Abdul Waheed, Hanin Atwany, Rita Singh, and Bhiksha Raj. 2025. On the robust approximation of asr metrics. *arXiv preprint arXiv:2502.12408*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, and 1 others. 2024. Llm inference unveiled: Survey and roofline model insights. *CoRR*.

Kamer Ali Yuksel, Thiago Castro Ferreira, Golara Javadi, Mohamed Al-Badrashiny, and Ahmet Gunduz. Norefer: a referenceless quality metric for automatic speech recognition via semi-supervised language model fine-tuning with contrastive learning.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.