Qwen3-ASR Technical Report

Qwen Team

https://huggingface.co/collections/Qwen/qwen3-asr
https://modelscope.cn/collections/Qwen/Qwen3-ASR
https://github.com/QwenLM/Qwen3-ASR

Abstract

In this report, we introduce Qwen3-ASR family, which includes two powerful all-in-one speech recognition models and a novel non-autoregressive speech forced alignment model. *Qwen3-ASR-1.7B* and *Qwen3-ASR-0.6B* are ASR models that support language identification and ASR for 52 languages and dialects. Both of them leverage large-scale speech training data and the strong audio understanding ability of their foundation model Qwen3-Omni. We conduct comprehensive internal evaluation besides the open-sourced benchmarks as ASR models might differ little on open-sourced benchmark scores but exhibit significant quality differences in real-world scenarios. The experiments reveal that the 1.7B version achieves state-of-the-art performance among open-sourced ASR models and is competitive with the strongest proprietary APIs while the 0.6B version offers the best accuracy–efficiency trade-off. Qwen3-ASR-0.6B can achieve an average time-to-first-token as low as 92ms and transcribe 2,000 seconds speech in 1 second at a concurrency of 128. *Qwen3-ForcedAligner-0.6B* is an LLM based NAR timestamp predictor that is able to align text-speech pairs in 11 languages. Timestamp accuracy experiments show that the proposed model outperforms the three strongest force alignment models and takes more advantages in efficiency and versatility. To further accelerate the community research of ASR and audio understanding, we release these models under the Apache 2.0 license.

1 Introduction

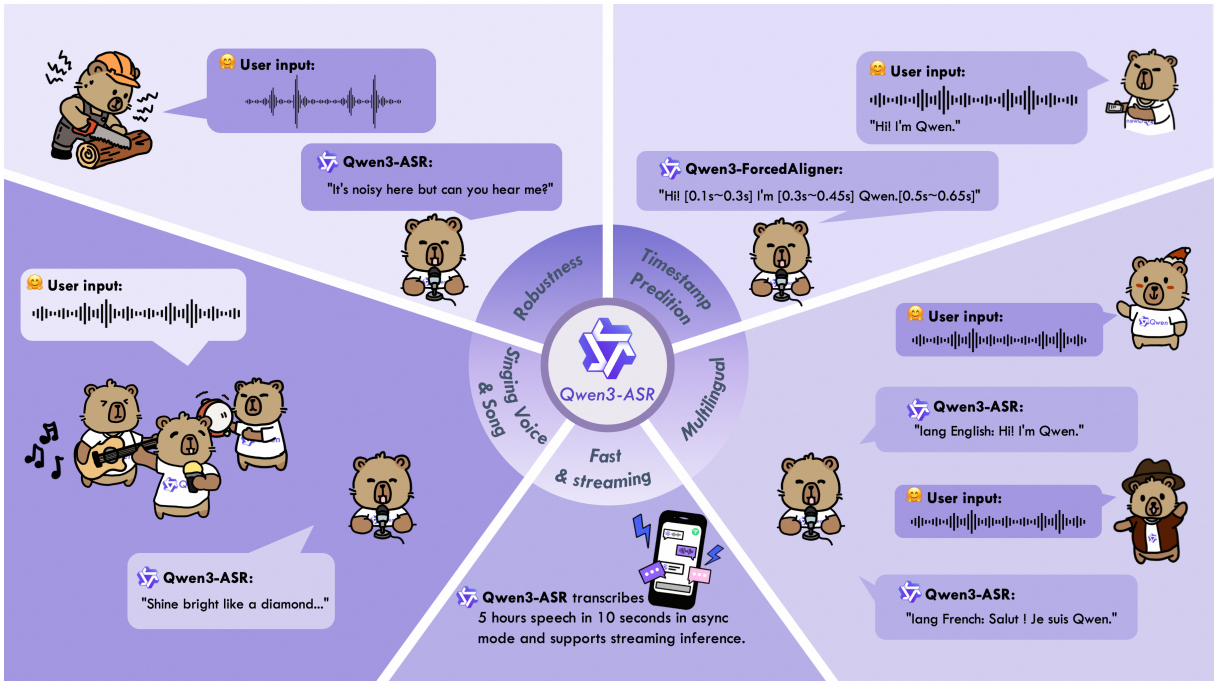

Figure 1: Qwen3-ASR family includes all-in-one ASR models with advantages in multilingual, noisy speech recognition, singing voice recognition and inference efficiency, so as to a novel multilingual speech forced alignment model for predicting timestamps of words or sentences in ASR results.

In recent years, Automatic Speech Recognition (ASR) has transitioned from traditional end-to-end (E2E) paradigms, e.g., Transducer (Graves, 2012) and AED (Chan et al., 2016; Radford et al., 2023) to the

Large Audio-Language Model (LALM) paradigm. Compared with traditional ASR, this paradigm can take advantage of the language modeling capabilities and world knowledge of large language models. The model first forms a high-level understanding of the audio signal and then generates transcription conditioned on this understanding, rather than relying solely on bottom-up acoustic pattern matching. Under this paradigm, issues that are relatively challenging for conventional ASR models—such as long-form transcription, robustness to noise, world-knowledge and named-entity recognition, as well as multilingual and dialectal coverage, can be addressed more naturally.

In real-world deployments, ASR systems are often required to output timestamps alongside transcripts (e.g., for subtitle generation). Prior work typically performs timestamping as a post-processing step using techniques such as CTC or CIF (Kürzinger et al., 2020; Rastorgueva et al., 2023; Shi et al., 2023). We would like to highlight that an LALM-based approach can yield more accurate and faster timestamp prediction at arbitrary temporal granularities, and that, by leveraging the multilingual capacity of LALMs, a single unified model can provide timestamp alignment across diverse languages.

In this report, we present the Qwen3-ASR family, including *Qwen3-ASR-1.7B* and *Qwen3-ASR-0.6B* - two all-in-one ASR models with language identification (LID) ability for 52 languages and dialects, and *Qwen3-ForcedAligner-0.6B* - the first lightweight LALM-based multilingual forced aligner supporting 11 languages and flexible timestamp prediction granularities. These models are posttrained from the strong foundation model of Qwen3-Omni (Xu et al., 2025a). For evaluating the performance of ASR models on benchmarks out of the open-sourced ones (ASR models at present have reached the limit of annotation errors on several test sets), we build a series of internal benchmarks covering more than complex acoustic environment, dialects, elders and kids speech and multilingual. Qwen3-ASR-1.7B achieves state-of-the-art (SOTA) performance among open-sourced ASR models and is competitive with the strongest proprietary commercial APIs. Qwen3-ASR-0.6B offers the best accuracy-model-size trade-off, making it a strong choice for on-device deployment. Qwen3-ForcedAligner-0.6B delivers highly accurate forced-alignment timestamps and inherits the key capabilities of Qwen3-ASR, including multilingual and long-form speech support, enabling scalable labeling of speech-transcript pairs.

The key features and contributions of the proposed Qwen3-ASR family models can be summarized as:

- **Achieves state-of-the-art all-in-one ASR and LID performance.** Qwen3-ASR-1.7B and Qwen3-ASR-0.6B finely support 30 languages, 22 Chinese dialects ASR, and English from countries and regions worldwide. These two models also conduct robust speech recognition under complex environment, including but not limited to singing voice and song recognition, noise environment recognition and complex text patterns recognition.
- **Presents novel speech force alignment architecture.** To the best of our knowledge, we introduce the first Large Language Model based speech forced aligner that produces accurate timestamps at flexible granularities, including word, sentence, and paragraph levels. In contrast to existing tools such as the Montreal Forced Aligner (MFA) and NeMo Forced Aligner (NFA), our model, Qwen3-ForcedAligner-0.6B, offers a unified, multilingual solution that addresses the lack of an all-in-one forced alignment system within the Qwen3-ASR family and fulfills a critical functional component for comprehensive spoken language processing.
- **Open-source models and a comprehensive inference and fine-tuning framework.** In addition to releasing the weights of three models, we provide a fully open-source, user-friendly codebase that supports inference with multiple features (e.g., multi-granularity alignment, streaming transcription, and multilingual processing) as well as a reproducible fine-tuning recipe. We hope this unified toolkit will accelerate research and development efforts in the automatic speech recognition community.

2 Qwen3-ASR

2.1 Architecture

Qwen3-ASR family models leverage Qwen3-Omni as a foundation model, which proved to obtain strong audio understanding ability (Xu et al., 2025b). Speech to recognize is first fed to AuT encoder, which is pretrained separately from Qwen3-Omni and Qwen3-ASR. As illustrated in Figure 2(left), AuT is an attention-encoder-decoder (AED) based ASR model which conducts 8 times downsampling to Fbank feature with 128 dimensions, yielding a 12.5Hz token rate audio encoder. We use a dynamic flash attention window size ranging from 1s to 8s, which allows Qwen3-ASR to perform both streaming inference with short chunks and offline inference with long queries. The architecture of models we are releasing is illustrated as Figure 2(right) and detailed as below: **Qwen3-ASR-1.7B** is built with Qwen3-1.7B, a projector, and an AuT encoder with 300M parameters and 1024 hidden size. This model demonstrates strong performance in multilingual and dialect speech recognition, so as well as robustness

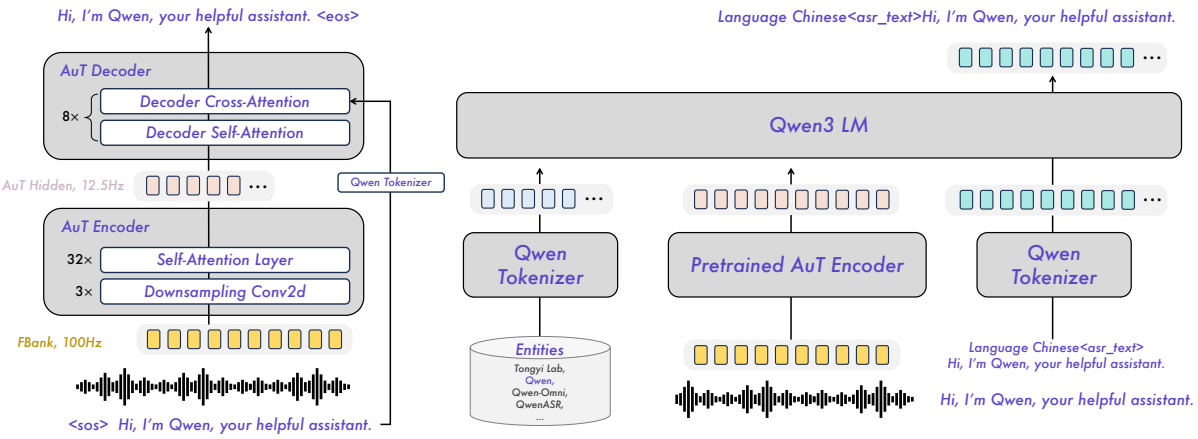

Figure 2: Architecture of AuT (left) and the overview of Qwen3-ASR (right).

under complex acoustic environment and text patterns. **Qwen3-ASR-0.6B** is built with Qwen3-0.6B, a projector and an AuT encoder with 180M parameters and a hidden size of 896. We design this compact model to balance recognition accuracy and inference efficiency, while remaining highly competitive among sub-1B-parameter ASR models.

2.2 Training Strategies

The training process of Qwen3-ASR consists of AuT pretraining, Omni pretraining, and ASR post-training, where the first two stages are identical to those of Qwen3-Omni.

(1) **AuT pretraining.** In this stage, we aim to obtain a pretrained encoder under the AED framework using large-scale labeled data. We leverage approximately 40 million hours of pseudo-labeled ASR data, where the majority is in Chinese and English. This pretrained encoder is shown to provide general and stable audio representations under dynamic attention window sizes.

(2) **Omni pretraining.** We use the pretrained Qwen3-Omni model as the foundation model for ASR training. Omni pretraining is conducted on multi-task audio, vision, and text data. In this stage, both Qwen3-ASR-0.6B and Qwen3-ASR-1.7B are trained with 3 trillion tokens, acquiring multi-modal understanding capability. The detailed training pipeline follows (Xu et al., 2025b).

(3) **ASR supervised finetuning (SFT).** In the SFT stage, we perform style transfer on the ASR input/output format with a substantially smaller set of multilingual data that is disjoint from the pretraining corpus. Besides standard Chinese, English and multilingual ASR data, SFT stage also utilizes non-speech data, streaming-enhancement data and context biasing data. Specifically, we train the model to be an ASR-only model that does not follow natural-language instructions in the prompt, in order to mitigate instruction injection and instruction-following failures. The output of Qwen3-ASR has outputs in two types for the given audio, with and without recognizable human speech:

Qwen3-ASR output style

Output style 1: For recognizable speech
<|**im_start**|>assistant
language English**<asr_text>**Today we release models including Qwen3-ASR-1.7B.**<|im_end|>**

Output style 2: For no speech detected
<|**im_start**|>assistant
language None**<asr_text><|im_end|>**

Meanwhile, the model learns to utilize the context tokens inside the system prompt as background knowledge, allowing users to obtain customized ASR results.

(4) **ASR reinforcement learning (RL).** At the last stage, we use Group Sequence Policy Optimization (GSPO, Zheng et al. (2025)) for further improving the quality of recognition. It turns out that RL plays an essential role in models' noise robustness, transcription stability and ability to analyze difficult cases. The total data leveraged by RL stage is about 50k utterances including 35% Chinese and English data, 35% multilingual data and 30% functional data, which aims at improving transcribing stability in complex environments.

2.3 Features

With the architecture and training strategies introduced above, Qwen3-ASR family models are notable in the aspects below:

Table 1: **Features of the Qwen3-ASR model family. Qwen3-ASR-1.7B and Qwen3-ASR-0.6B support 52 languages and dialects, comprising 30 languages and 22 Chinese dialects. Qwen3-ForcedAligner-0.6B supports 11 languages. Seq. Len. denotes the maximum audio length for single inference in seconds, and NAR denotes non-autoregressive inference.**

Model	Supported Languages	Supported Dialects	Inference Mode	Seq. Len.	Audio Types
Qwen3-ASR-1.7B & **Qwen3-ASR-0.6B**	Chinese (zh), English (en), Cantonese (yue), Arabic (ar), German (de), French (fr), Spanish (es), Portuguese (pt), Indonesian (id), Italian (it), Korean (ko), Russian (ru), Thai (th), Vietnamese (vi), Japanese (ja), Turkish (tr), Hindi (hi), Malay (ms), Dutch (nl), Swedish (sv), Danish (da), Finnish (fi), Polish (pl), Czech (cs), Filipino (fil), Persian (fa), Greek (el), Hungarian (hu), Macedonian (mk), Romanian (ro)	Anhui, Dongbei, Fujian, Gansu, Guizhou, Hebei, Henan, Hubei, Hunan, Jiangxi, Ningxia, Shandong, Shaanxi, Shanxi, Sichuan, Tianjin, Yunnan, Zhejiang. Cantonese (Hong Kong accent), Cantonese (Guangdong accent), Wu language, Minnan language.	Offline / Streaming	1200s	Speech, Singing Voice, Songs with BGM
Qwen3-ForcedAligner-0.6B	Chinese, English, Cantonese, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish	–	NAR	300s	Speech

(1) **Accurate Chinese and English ASR.** Chinese and English account for the majority of the training data across all stages, and the model achieves leading Chinese and English recognition performance over multiple benchmarks compared with many competing systems.

(2) **Multilingual, multiple dialects supporting.** Qwen3-ASR-1.7B and Qwen3-ASR-0.6B support 30 languages and 22 dialects, detailed in Table 1.

(3) **Long-form and streaming inference.** Qwen3-ASR-1.7B and Qwen3-ASR-0.6B naturally support single speech no longer than 20 minutes and streaming/offline unified inference.

(4) **Singing voice and songs recognition.** Qwen3-ASR-1.7B and Qwen3-ASR-0.6B recognize singing voice and songs accurately. In addition to achieving strong singing-voice recognition, the Qwen3-ASR family also supports direct transcription of complete songs with background music (BGM), demonstrating robustness to accompaniment and complex musical mixtures.

2.4 Inference Efficiency

The speed benchmarks of Qwen3-ASR are conducted in two settings: offline batch inference and online asynchronous inference. The former is evaluated using vLLM's offline batch generation, while the latter is evaluated with a multi-concurrency request setup based on vLLM Serve, which better reflects inference efficiency in industrial environments. All experiments are run with vLLM v0.14.0, with CUDA Graph enabled and bfloat16 precision for inference. The results in Table 2 show that, under different concurrency levels, Qwen3-ASR-0.6B can achieve an average Time-to-First-Token (TTFT) as low as 92ms. It reaches a real-time factor (RTF) as low as 0.064 and throughput as high as 2000 at a concurrency of 128, which means it can process 2,000 seconds of audio per second.

3 Qwen3-ForcedAligner

3.1 Overview

Qwen3-ForcedAligner-0.6B aims to estimate the start and end timestamps of each word or character in a speech, given the corresponding transcript. Qwen3-ForcedAligner-0.6B reframes the forced alignment task within a slot-filling formulation. Specifically, given a speech and a transcript augmented with special tokens *[time]* that denote word-level or character-level start and end timestamp slots, Qwen3-ForcedAligner-0.6B directly predicts the corresponding discrete timestamp indices for each slot.

The key features and contributions of Qwen3-ForcedAligner-0.6B can be summarized as:

Table 2: **Efficiency of Qwen3-ASR family models. Qwen3-ASR-0.6B and Qwen3-ASR-1.7B support vLLM-based inference in both offline batch and online asynchronous mode, while Qwen3-ForcedAligner-0.6B supports offline batch inference in PyTorch only. All measurements in the table are based on input audio of approximately 2 minutes for ASR and 1 minute for FA in length, and all inference is performed on a single typical computing resource. Conc. denotes the concurrency level. TTFT p95 denotes the 95th percentile TTFT latency.**

Model		Offline		Online async			
	Conc.	RTF	Throughput	TTFT avg. (ms)	TTFT p95 (ms)	RTF	Throughput
Qwen3-ASR-0.6B	1	0.00923	108.34	92	105	0.00940	106.38
	2	0.01124	177.94	103	168	0.01108	180.51
	4	0.01284	311.53	132	203	0.01224	326.80
	8	0.01600	500.00	228	417	0.01472	543.48
	16	0.02384	671.14	459	882	0.01936	826.45
	32	0.03808	840.34	820	1575	0.02912	1098.90
	64	0.06336	1010.10	1631	3196	0.04352	1470.59
	128	0.11264	1136.36	3210	6195	0.06400	2000.00
	256	0.21504	1190.48	-	-	-	-
	512	0.44544	1149.43	-	-	-	-
Qwen3-ASR-1.7B	1	0.01482	67.48	102	113	0.01483	67.43
	2	0.01540	129.87	117	170	0.01530	130.72
	4	0.01712	233.64	135	192	0.01688	236.97
	8	0.02072	386.10	224	382	0.02000	400.00
	16	0.02896	552.49	443	791	0.02640	606.06
	32	0.04608	694.44	847	1570	0.03968	806.45
	64	0.07360	869.57	1597	2942	0.06208	1030.93
	128	0.13056	980.39	3392	6227	0.10496	1219.51
	256	0.24320	1052.63	-	-	-	-
	512	0.50176	1020.41	-	-	-	-
Qwen3-ForcedAligner-0.6B	1	0.00889	112.49	-	-	-	-
	2	0.00232	862.07	-	-	-	-
	4	0.00432	925.93	-	-	-	-
	8	0.00832	961.54	-	-	-	-
	16	0.01696	943.40	-	-	-	-
	32	0.03584	892.86	-	-	-	-
	64	0.08192	781.25	-	-	-	-
	128	0.19712	649.35	-	-	-	-

- **Accurate Timestamp Prediction.** Qwen3-ForcedAligner-0.6B exhibits substantially lower timestamp prediction shifts, achieving a relative reduction of 67%~77% in accumulated average shift on the human-labeled test datasets compared with other forced alignment methods.
- **Broad Application Scenarios.** Qwen3-ForcedAligner-0.6B supports speech in 11 languages with durations of up to 300 seconds, including cross-lingual scenarios, and allows users to flexibly customize timestamp prediction for any word or character.
- **Fast Inference Speed.** Qwen3-ForcedAligner-0.6B abandons the next-token prediction paradigm and adopts non-autoregressive (NAR) inference for timestamp prediction.

3.2 Model Design

As shown in Figure 3, Qwen3-ForcedAligner-0.6B employs a pretrained AuT encoder to process the input speech signal and obtain speech embeddings. The transcript is reformatted by appending start and end timestamp labels to each word or character, after which each timestamp label is replaced with a special token *[time]* and fed into the tokenizer. Moreover, the timestamp labels in the transcript are discretized into indices by dividing each timestamp value by the 80ms frame duration of the AuT encoder output. Speech and text embedding sequences are processed by the Qwen3-0.6B LLM, followed by a timestamp prediction linear layer that predicts timestamp indices for the entire input sequence. In this work, the maximum number of classes is 3,750, corresponding to support for speech inputs of up to 300s.

The AuT encoder and the multilingual Qwen3-0.6B LLM jointly provide Qwen3-ForcedAligner-0.6B with multilingual and cross-lingual capabilities. Specifically, the AuT encoder, pretrained on a large-scale multilingual corpus, generates effective frame-level speech embeddings for multiple languages, while the multilingual Qwen3-0.6B LLM handles semantic information across different languages. In addition, the special token *[time]* and the timestamp prediction layer do not rely on language-specific phoneme sets or dictionaries. Details can be found in Mu et al. (2026b).

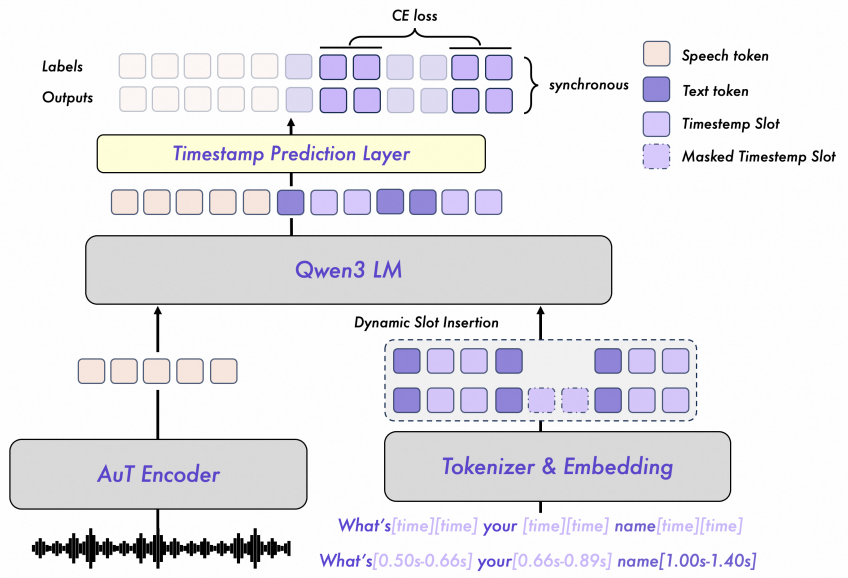

Figure 3: Illustration of Qwen3-ForcedAligner-0.6B. During training, randomly masked timestamp slots are dynamically inserted into the token sequence to represent word or character boundaries. The combined sequence is fed into Qwen3-0.6B LLM, and a timestamp prediction layer predicts the corresponding timestamp indices for each slot. Supervision is applied with cross-entropy loss on synchronously aligned label and output sequences.

3.3 Training Strategies

Training Qwen3-ForcedAligner-0.6B requires word-level or character-level timestamp labels for a large number of speech–transcript pairs. However, because manual annotation is prohibitively expensive, we use pseudo-timestamp labels generated by the Montreal forced aligner (MFA) McAuliffe et al. (2017), which is among the most accurate existing forced alignment methods. It is important to note that MFA pseudo-labels inherently contain noise and systematic shifts. Qwen3-ForcedAligner does not simply replicate MFA outputs; instead, it distills and smooths these pseudo-labels, resulting in more stable timestamp predictions with reduced shift.

LALMs typically use a training scheme in which the last token of the output sequence and the first token of the label sequence are removed, creating a one-position offset between the two sequences; the cross-entropy loss is then computed to implement the standard next-token prediction paradigm. However, this paradigm is not suitable for filling timestamp slots. Qwen3-ForcedAligner-0.6B employs causal training, keeping the output and label sequences non-shifted, which allows the model to explicitly recognize timestamp slots during training and predict the timestamp indices to fill them. Moreover, causal training enables Qwen3-ForcedAligner-0.6B to incorporate prior contextual information when predicting the timestamp for the current slot, ensuring global consistency in timestamp prediction. The cross-entropy loss is computed only in the timestamp slots, thereby focusing the training objective of Qwen3-ForcedAligner-0.6B on timestamp slot filling.

In addition, Qwen3-ForcedAligner-0.6B employs a dynamic slot insertion strategy during training to enhance its generalization capability. Specifically, for each word or character in a sample, the model randomly determines whether to insert start and end timestamp slots afterward.

3.4 Inference and Usability

Since the token sequences remain non-shifted during training, users can insert start and end timestamp slots after any word or character, and Qwen3-ForcedAligner-0.6B uses non-autoregressive (NAR) decoding to predict the timestamp indices for all slots in the transcript simultaneously. Once the timestamp indices are obtained, multiplying each index by 80ms recovers the actual predicted timestamps.

The speed benchmark for Qwen3-ForcedAligner is conducted with FlashAttention and bfloat16. Since the model is non-autoregressive, the inference speed difference between Transformers and vLLM is relatively small; therefore, all our benchmarks are run with Transformers. The results in Table 2 show that the model can maintain an RTF close to 0.001 even under high concurrency, i.e., it can process 1,000 seconds of audio per second.

4 Experiments

4.1 Evaluation Details

Baseline Systems. To validate the Qwen3-ASR family, we conduct comparative evaluations against state-of-the-art (SOTA) closed-source ASR APIs and widely used open-source models. Specifically, we compare Qwen3-ASR with three leading proprietary services: GPT-4o-Transcribe (OpenAI, 2024), Gemini-2.5-Pro (Comanici et al., 2025), and Doubao-ASR (Bai et al., 2024). We further include several multilingual open-source baselines, namely Whisper-large-v3 (Radford et al., 2023), FunASR-MLT-Nano (An et al., 2025), and GLM-ASR-Nano (Z.ai, 2025). Together, these baselines represent strong commercial systems and competitive open-source alternatives, enabling a comprehensive evaluation of Qwen3-ASR under representative real-world conditions.

Benchmark Introduction. We adopt a four-part evaluation protocol to measure the speech recognition performance of the proposed Qwen3-ASR series:

1. **Public benchmarks (English and Chinese).** We evaluate a broad set of public benchmarks (Conneau et al., 2023; Ardila et al., 2020; Zhang et al., 2022; Panayotov et al., 2015; Dai et al., 2025; Li et al., 2025) and report the results separately for subsets of English, standard Mandarin and Chinese dialects, including two recently released benchmarks.
2. **Internal robustness suite.** We stress-test the model under challenging real-world conditions using a comprehensive in-house suite, covering English speech from multiple countries and accents (16 accent groups in total), 22 Chinese dialect varieties, and difficult scenarios including elderly and children's speech, extremely low signal-to-noise (SNR) ratios, nonfluent and tongue-twister-like repetitive speech, and multi-speaker Chinese conversational speech. These settings enable a systematic assessment of robustness to accent/dialect variability and complex acoustic and linguistic conditions.
3. **Multilingual evaluation.** The model supports ASR for 30 languages. We evaluate on Common Voice, Fleurs, MLS, MLC-SLM (Mu et al., 2026a), and an internally curated test set spanning 15 languages. The language inventory of each benchmark is specified in Section 2.3. Since Fleurs covers a particularly large and diverse set of languages, we additionally report results on progressively expanded language subsets grouped by language popularity and practical usage for a more fine-grained characterization. Meaning while, we evaluate the language identification performance on the multilingual open-source benchmarks.
4. **Singing voice recognition.** We evaluate singing voice transcription on both public benchmarks and an internal test set. In the internal evaluation, we emphasize long-form transcription where an entire song is provided as a single input, to assess robustness to long-duration audio as well as the distinctive acoustic and rhythmic properties of singing.

Evaluation Metrics. For recognition accuracy, we report either **word error rate (WER)** or **character error rate (CER)** depending on the language. We use CER for character-based languages (e.g., Mandarin Chinese, Cantonese, and Korean) and WER for word-delimited languages (e.g., English, German, and French). When aggregated results are needed (e.g., average performance across multiple languages or dialects), we report the macro-average (i.e., the unweighted mean across languages/dialects). The best result in each table is highlighted in **bold**. In addition, when Qwen3-ASR-0.6B is the best-performing model after excluding the larger Qwen3-ASR-1.7B, we also highlight it in bold.

For language identification, we report **language identification accuracy**.

For timestamp accuracy, Qwen3-ForcedAligner uses **Accumulated Average Shift** (AAS Shi et al. (2023)), where lower values indicate more accurate timestamp predictions. AAS is defined as the mean absolute difference between predicted timestamps and reference timestamps over all timestamp slots in the evaluated datasets:

$$\mathrm{AAS} = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{n}_i - n_i\right|, \tag{1}$$

where N is the total number of timestamp slots, $\hat{n}_i$ denotes the timestamp predicted by Qwen3-ForcedAligner for slot i, and n_i is the corresponding reference timestamp obtained from Montreal Forced Aligner (MFA) or manual annotations.

4.2 English & Chinese ASR Performance

4.2.1 Opensource ASR Benchmarks

As shown in Table 3, Qwen3-ASR delivers consistently strong performance across English, Mandarin Chinese, and multiple Chinese dialect benchmarks. It is competitive with leading commercial APIs

while substantially outperforming widely used open-source baselines. Scaling from Qwen3-ASR-0.6B to Qwen3-ASR-1.7B yields clear and stable gains, indicating that the model benefits effectively from increased capacity.

On **English** benchmarks, Qwen3-ASR performs particularly well on diverse, real-world data (e.g., crowd-sourced or web-collected speech), where distribution shift is more pronounced than in read-speech settings. In these cases, Qwen3-ASR-1.7B achieves the strongest overall results on several datasets, while remaining close to the best-performing systems on standard academic evaluations such as LibriSpeech. Compared with commercial APIs, whose performance can vary substantially across datasets, Qwen3-ASR shows more consistent accuracy across a broad range of English conditions.

On **Mandarin Chinese**, Qwen3-ASR demonstrates a clear advantage. It delivers the best overall performance on most Mandarin benchmarks in the table and remains reliable on more challenging large-scale evaluations. Notably, on WenetSpeech, which contains diverse acoustic environments and meeting-style speech, Qwen3-ASR outperforms the available baselines by a large margin.

On **Chinese dialect** benchmarks, Qwen3-ASR maintains strong accuracy under substantial pronunciation and lexical variation. It consistently ranks among the top systems across Cantonese and other dialect datasets, and performs particularly well on more challenging long-utterance settings, demonstrating robustness beyond short, clean test conditions. While a small number of dialect-specific cases favor specialized commercial APIs, Qwen3-ASR remains highly competitive overall and provides a strong general-purpose solution across dialects without per-dialect customization.

Overall, Table 3 highlights three key advantages of Qwen3-ASR: (i) strong cross-domain generalization on English benchmarks beyond curated read speech, (ii) state-of-the-art accuracy on Mandarin Chinese across multiple public datasets including large-scale, noisy meeting-style speech, and (iii) robust handling of Chinese dialects, with especially strong performance on Cantonese and long/short dialectal speech. These findings demonstrate that Qwen3-ASR delivers strong, reproducible performance across diverse public benchmarks, while also remaining competitive with top-tier closed-source APIs.

Table 3: **Evaluation on English, Mandarin Chinese, and a range of Chinese dialect benchmarks. For the commercial APIs and the open-source Whisper-large-v3 model, we obtained results by running inference on the test sets ourselves due to the absence of published numbers; for FunASR-MLT-Nano, we report the results from its official technical report. "N/A" denotes that we cannot get a reasonable result by the official API. "–" indicates that the corresponding benchmark result is not reported.**

	GPT-4o -Transcribe	Gemini-2.5 -Pro	Doubao-ASR	Whisper -large-v3	Fun-ASR -MLT-Nano	Qwen3-ASR -0.6B	Qwen3-ASR -1.7B
English (en)							
LibriSpeech *clean* | *other*	**1.39** | 3.75	2.89 | 3.56	2.78 | 5.70	1.51 | 3.97	1.68 | 4.03	2.11 | 4.55	1.63 | **3.38**
GigaSpeech	25.50	9.37	9.55	9.76	-	**8.88**	**8.45**
CV-en	9.08	14.49	13.78	9.90	9.90	9.92	**7.39**
Fleurs-en	**2.40**	2.94	6.31	4.08	5.49	4.39	3.35
MLS-en	5.12	**3.68**	7.09	4.87	-	6.00	4.58
Tedlium	7.69	6.15	4.91	6.84	-	**3.85**	**4.50**
VoxPopuli	10.29	11.36	12.12	12.05	-	**9.96**	**9.15**
Chinese (zh)							
WenetSpeech *net* | *meeting*	15.30 | 32.27	14.43 | 13.47	N/A	9.86 | 19.11	6.35 | -	**5.97** | **6.88**	**4.97** | **5.88**
AISHELL-2-test	4.24	11.62	2.85	5.06	-	3.15	**2.71**
SpeechIO	12.86	5.30	2.93	7.56	-	3.44	**2.88**
Fleurs-zh	2.44	2.71	2.69	4.09	3.51	2.88	**2.41**
CV-zh	6.32	7.70	5.95	12.91	6.20	6.89	**5.35**
Chinese Dialect							
KeSpeech	26.87	24.71	5.27	28.79	-	7.08	**5.10**
Fleurs-yue	4.98	9.43	4.98	9.18	-	5.79	**3.98**
CV-yue	11.36	18.76	13.20	16.23	-	9.50	**7.57**
CV-zh-tw	6.32	7.31	4.06	7.84	-	5.59	**3.77**
WenetSpeech-Yue *short* | *long*	15.62 | 25.29	25.19 | 11.23	9.74 | 11.40	32.26 | 46.64	- | -	7.54 | 9.92	**5.82** | **8.85**
WenetSpeech-Chuan *easy* | *hard*	34.81 | 53.98	43.79 | 67.30	**11.40** | **20.20**	14.35 | 26.80	- | -	13.92 | 24.45	11.99 | 21.63

4.2.2 Internal ASR Benchmarks

To further assess robustness in realistic deployment settings, we evaluate Qwen3-ASR on our internal robustness suite; results are summarized in Table 4. Qwen3-ASR delivers consistently strong performance across all subsets, and scaling from 0.6B to 1.7B yields stable gains. In the accented-English evaluation,

Table 4: **Evaluation on internal English and Chinese test sets covering multiple accents and dialects, as well as challenging acoustic conditions and difficult speaking scenarios.**

	GPT-4o -Transcribe	Gemini-2.5 -Pro	Doubao-ASR	Whisper -large-v3	Fun-ASR -MLT-Nano	Qwen3-ASR -0.6B	Qwen3-ASR -1.7B
Accented English							
Dialog-Accented English	28.56	23.85	20.41	21.30	19.96	**16.62**	**16.07**
Chinese Mandarin							
Elders&Kids	14.27	36.93	4.17	10.61	4.54	4.48	**3.81**
ExtremeNoise	36.11	29.06	17.04	63.17	36.55	17.88	**16.17**
TongueTwister	20.87	4.97	3.47	16.63	9.02	4.06	**2.44**
Dialog-Mandarin	20.73	12.50	6.61	14.01	7.32	7.06	**6.54**
Chinese Dialect							
Dialog-Cantonese	16.05	14.98	7.56	31.04	5.85	**4.80**	**4.12**
Dialog-Chinese Dialects	45.37	47.70	19.85	44.55	19.41	**18.24**	**15.94**

Dialect coverage: Results for *Dialog-Accented English* are averaged over 16 accents, and results for *Dialog-Chinese Dialects* are averaged over 22 Chinese dialects. Detailed category definitions are provided in Section 2.3.

Qwen3-ASR achieves the lowest WER among all compared systems, surpassing both commercial APIs and open-source baselines, indicating better generalization to accent variation. On Mandarin, Qwen3-ASR-1.7B performs best across all evaluated subsets, demonstrating robustness under difficult acoustic and speaking conditions. In dialectal Chinese, Qwen3-ASR again achieves the best results on both conversational Cantonese and the aggregated 22-dialect evaluation; the gains are particularly pronounced in the multi-dialect mixture, highlighting improved robustness as linguistic diversity increases. Overall, these internal results are consistent with the public-benchmark findings and further confirm that Qwen3-ASR provides reliable recognition quality in high-variability scenarios.

4.3 Multilingual ASR and Language Identification

4.3.1 Multilingual ASR Performance

Table 5: **Evaluation of multilingual ASR systems on a comprehensive set of benchmark datasets.**

	GLM-ASR -Nano-2512	Whisper -large-v3	Fun-ASR -MLT-Nano	Qwen3-ASR -0.6B	Qwen3-ASR -1.7B
Open-sourced Benchmarks					
MLS	13.32	8.62	28.70	13.19	**8.55**
CommonVoice	19.40	10.77	17.25	12.75	**9.18**
MLC-SLM	34.93	15.68	29.94	15.84	**12.74**
Fleurs	16.08	5.27	10.03	7.57	**4.90**
Fleurs†	20.05	6.85	31.89	10.37	**6.62**
Fleurs††	24.83	**8.16**	47.84	21.80	12.60
Qwen-ASR Internal Benchmarks					
News-Multilingual	49.40	14.80	65.07	17.39	**12.80**

Language coverage: *MLS* includes 8 languages: {da, de, en, es, fr, it, pl, pt}.
CommonVoice includes 13 languages: {en, zh, yue, zh_TW, ar, de, es, fr, it, ja, ko, pt, ru}.
MLC-SLM includes 11 languages: {en, fr, de, it, pt, es, ja, ko, ru, th, vi}.
Fleurs includes 12 languages: {en, zh, yue, ar, de, es, fr, it, ja, ko, pt, ru }.
Fleurs† includes 8 additional languages beyond *Fleurs*: {hi, id, ms, nl, pl, th, tr, vi}.
Fleurs†† includes 10 additional languages beyond *Fleurs*†: {cs, da, el, fa, fi, fil, hu, mk, ro, sv}.
News-Multilingual includes 15 languages: {ar, de, es, fr, hi, id, it, ja, ko, nl, pl, pt, ru, th, vi}.

In this part, we illustrate the multilingual ASR performance of the Qwen3-ASR series on a broad set of public benchmarks as well as our internal multilingual news evaluation (Table 5). Overall, Qwen3-ASR-1.7B achieves the best average performance on most test settings, showing strong generalization across languages and domains, while Qwen3-ASR-0.6B provides a competitive lightweight alternative.

On MLS, Common Voice and MLC-SLM benchmarks, Qwen3-ASR-1.7B consistently outperforms the evaluated open-source baselines, including the widely used Whisper-large-v3, and substantially surpasses smaller multilingual models. For Fleurs, which spans more languages and diverse recording conditions, Qwen3-ASR-1.7B achieves the best performance on the 12- and 20-language subsets. However, relative to Whisper-large-v3, its performance degrades on the full 30-language setting, indicating room for improvement in handling increased linguistic diversity and long-tail languages. Nevertheless,

Qwen3-ASR-1.7B remains markedly better than the 0.6B variant, suggesting that model scaling improves robustness in more challenging multilingual regimes.

Finally, on our internal News-Multilingual benchmark, Qwen3-ASR-1.7B achieves the best overall performance, demonstrating stronger robustness to domain shift (e.g., broadcast/news-style speech) than all baselines. Overall, these results indicate effective scaling behavior and strong multilingual recognition across both public and internal evaluations. Per-language results for the Qwen3-ASR family are provided in the Appendix.

4.3.2 Language Identification Performance

Table 6: **Language identification accuracy (%) ↑ on open-source multilingual test sets.**

	Whisper-large-v3	Qwen3-ASR-0.6B	Qwen3-ASR-1.7B
MLS	**99.9**	99.3	**99.9**
CommonVoice	92.7	**98.2**	**98.7**
MLC-SLM	89.2	**92.7**	**94.1**
Fleurs	94.6	**97.1**	**98.7**
Avg.	94.1	**96.8**	**97.9**

Language coverage: The language sets follow Table 5. Here, Fleurs corresponds to Fleurs†† in Table 5 and covers 30 languages.

Following the output template in Section 2.2, Qwen3-ASR not only decodes speech into text, but also performs language identification (LID) via natural-language prompting before ASR decoding. In this section, we evaluate LID accuracy on 4 multilingual benchmarks: Fleurs (30 languages), MLS (9 languages), CommonVoice (13 languages), MLC-SLM (11 languages); the covered languages are detailed in Section 2.3. As shown in Table 6, we compare Qwen3-ASR-0.6B and Qwen3-ASR-1.7B with Whisper-large-v3, a strong multilingual ASR model with built-in LID capability. Both Qwen3-ASR models outperform Whisper-large-v3, demonstrating stable and effective language identification across these mainstream languages. Most remaining errors on Fleurs stem from confusion between Malay (ms) and Indonesian (id), two closely related languages with high acoustic similarity.

4.4 Singing Voice & Songs Recognition Performance

Table 7: **Singing-voice and song-transcription results. WER (%) is reported for singing-only benchmarks and long-form songs with background music. "N/A" indicates that the model does not support long-form song recognition due to the poor performance.**

	GPT-4o -Transcribe	Gemini-2.5 -Pro	Doubao-ASR -1.0	Whisper -large-v3	Fun-ASR-MLT -Nano	Qwen3-ASR -1.7B
Singing						
M4Singer	16.77	20.88	7.88	13.58	7.29	**5.98**
MIR-1k-vocal	11.87	9.85	6.56	11.71	8.17	**6.25**
Opencpop	7.93	6.49	3.80	9.52	**2.98**	3.08
Popcs	32.84	15.13	8.97	13.77	9.42	**8.52**
Songs with BGM						
EntireSongs-en	30.71	**12.18**	33.51	N/A	N/A	14.60
EntireSongs-zh	34.86	18.68	23.99	N/A	N/A	**13.91**

Table 7 reports results for singing-voice transcription and long-form song transcription with background music. Overall, Qwen3-ASR-1.7B is robust to melody-induced pronunciation variation and musical accompaniment, outperforming most commercial APIs and open-source baselines across the evaluated sets. For **singing-only** benchmarks, it achieves the best performance for M4Singer, MIR-1k-vocal, and Popcs, while remaining competitive for Opencpop (second to FunASR-MLT-Nano by a small margin), indicating strong generalization across singing styles and recording conditions with reduced sensitivity to pitch drift, phoneme elongation, and rhythmic lyric variation. For **full songs with background music**, Qwen3-ASR-1.7B substantially outperforms open-source baselines; Whisper-large-v3 and FunASR-MLT-Nano degrade markedly in long-form, music-mixed settings. It achieves high accuracy for both English and Chinese songs, ranking first on the Chinese set and remaining competitive with the best commercial system on the English set, suggesting that Qwen3-ASR is well suited to realistic music-containing scenarios and background-music-robust and narrows the gap between speech ASR and singing/song transcription.

4.5 Streaming Speech Recognition

This section evaluates Qwen3-ASR-1.7B and Qwen3-ASR-0.6B in both offline and streaming inference modes. Benefiting from the dynamic attention-window mechanism, the Qwen3-ASR family supports streaming inference naturally. Table 8 reports results on three open-source test sets using a 2-second chunk size, a 5-token fallback, and keeping the last four chunks unfixed. Overall, Qwen3-ASR provides a unified model for offline and streaming use, while streaming inference preserves strong recognition accuracy.

Table 8: **ASR performance of the two inference modes on three open-source benchmarks.**

Model	Infer. Mode	Librispeech	Fleurs-en	Fleurs-zh	Avg.
Qwen3-ASR-1.7B	Offline	1.63 | 3.38	3.35	2.41	2.69
	Streaming	1.95 | 4.51	4.02	2.84	3.33
Qwen3-ASR-0.6B	Offline	2.11 | 4.55	4.39	2.88	3.48
	Streaming	2.54 | 6.27	5.38	3.40	4.40

4.6 Precision of Timestamps

Table 9: **Accumulated Average Shift (AAS, ms) ↓ of Qwen3-ForcedAligner-0.6B and competing forced-alignment methods on MFA-labeled and human-labeled test sets.**

	Monotonic-Aligner	NFA	WhisperX	Qwen3-ForcedAligner-0.6B
MFA-Labeled Raw				
Chinese	161.1	109.8	-	**33.1**
English	-	107.5	92.1	**37.5**
French	-	100.7	145.3	**41.7**
German	-	122.7	165.1	**46.5**
Italian	-	142.7	155.5	**75.5**
Japanese	-	-	-	**42.4**
Korean	-	-	-	**37.2**
Portuguese	-	-	-	**38.4**
Russian	-	200.7	-	**40.2**
Spanish	-	124.7	108.0	**36.8**
Avg.	161.1	129.8	133.2	**42.9**
MFA-Labeled Concat-300s				
Chinese	1742.4	235.0	-	**36.5**
English	-	226.7	227.2	**58.6**
French	-	230.6	2052.2	**53.4**
German	-	220.3	993.4	**62.4**
Italian	-	290.5	5719.4	**81.6**
Japanese	-	-	-	**81.3**
Korean	-	-	-	**42.2**
Portuguese	-	-	-	**50.0**
Russian	-	283.3	-	**43.0**
Spanish	-	240.2	4549.9	**39.6**
Cross-lingual	-	-	-	**34.2**
Avg.	1742.4	246.7	2708.4	**52.9**
Human-Labeled				
Raw	49.9	88.6	-	**27.8**
Raw-Noisy	53.3	89.5	-	**41.8**
Concat-60s	51.1	86.7	-	**25.3**
Concat-300s	410.8	140.0	-	**24.8**
Concat-Cross-lingual	-	-	-	**42.5**
Avg.	141.3	101.2	-	**32.4**

Table 9 reports the AAS of Qwen3-ForcedAligner-0.6B and competing forced-alignment methods on MFA-labeled and human-labeled test sets. Competing methods require language-specific models and support only a limited set of languages, whereas Qwen3-ForcedAligner-0.6B covers multiple languages with a single model and supports cross-lingual, code-switched scenarios. In addition, Qwen3-ForcedAligner-0.6B performs consistently on both short and long utterances, while baseline methods show a sharp degradation in timestamp accuracy on long utterances. Although trained with MFA pseudo-labels,

Qwen3-ForcedAligner-0.6B still achieves low AAS on the human-labeled test sets, indicating strong real-world generalization.

5 Conclusion

We present Qwen3-ASR, a model family comprising two automatic speech recognition (ASR) systems and a forced-alignment (FA) model trained on large-scale speech corpora. By leveraging the strong audio understanding capability of the foundation model Qwen3-Omni and a four-stage training pipeline, Qwen3-ASR-1.7B and Qwen3-ASR-0.6B outperform competing models of comparable or larger size, as well as commercial APIs, in both speech coverage and recognition accuracy. The models support language identification and ASR across 30 languages, deliver robust performance in complex acoustic conditions, exhibit resilience to accents and dialects, and maintain effectiveness on singing voice and other real-world speech scenarios. In addition, we introduce Qwen3-ForcedAligner-0.6B, an LLM-based non-autoregressive timestamp predictor that enables forced alignment for 11 languages with end-to-end processing times under five minutes. This approach surpasses three mainstream end-to-end ASR-based FA solutions in timestamp accuracy, inference speed, and language coverage. Alongside releasing the weights for all three models, we open-source a unified and user-friendly inference framework. Overall, the Qwen3-ASR family achieves state-of-the-art performance on real-world evaluations and public benchmarks, and the open-sourced forced-alignment model addresses a critical gap in the speech technology stack. We will continue to advance this open model family in accuracy and functionality.

6 Authors

Core Contributors: Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu†, Jingren Zhou, Junyang Lin†

Contributors[1]: Yunfei Chu, Daren Chen, Ting He, Hangrui Hu, Jiayi Leng, Zheng Li, Yuanjun Lv, Bingshen Mu, Hao Su, Xian Yang, Xuechun Wang, Yuezhang Wang, Zhenglin Wang, Lei Xie, Jianwei Zhang, Xinfa Zhu, Guangdong Zhou

References

Keyu An, Yanni Chen, Zhigao Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, Ying Liu, Xiang Lv, Yunjie Ji, Yiheng Jiang, Bin Ma, Haoneng Luo, Chongjia Ni, Zexu Pan, Yiping Peng, Zhendong Peng, Peiyao Wang, Hao Wang, Haoxu Wang, Wen Wang, Wupeng Wang, Yuzhong Wu, Biao Tian, Zhentao Tan, Nan Yang, Bin Yuan, Jieping Ye, Jixing Yu, Qinglin Zhang, Kun Zou, Han Zhao, Shengkui Zhao, Jingren Zhou, and Yanqiao Zhu. Fun-ASR Technical Report, 2025. URL https://arxiv.org/abs/2509.12508.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4218–4222, 2020.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, Lu Gao, Yi Guo, Minglun Han, Ting Han, Wenchao Hu, Xinying Hu, Yuxiang Hu, Deyu Hua, Lu Huang, Mingkun Huang, Youjia Huang, Jishuo Jin, Fanliu Kong, Zongwei Lan, Tianyu Li, Xiaoyang Li, Zeyang Li, Zehua Lin, Rui Liu, Shouda Liu, Lu Lu, Yizhou Lu, Jingting Ma, Shengtao Ma, Yulin Pei, Chen Shen, Tian Tan, Xiaogang Tian, Ming Tu, Bo Wang, Hao Wang, Yuping Wang, Yuxuan Wang, Hanzhang Xia, Rui Xia, Shuangyi Xie, Hongmin Xu, Meng Yang, Bihong Zhang, Jun Zhang, Wanyi Zhang, Yang Zhang, Yawei Zhang, Yijie Zheng, and Ming Zou. Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition, 2024. URL https://arxiv.org/abs/2407.04675.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. ICASSP*, pp. 4960–4964. IEEE, 2016. URL https://doi.org/10.1109/ICASSP.2016.7472621.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

[1]Alphabetical order. †Corresponding Authors.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. IEEE, 2023.

Yuhang Dai, Ziyu Zhang, Shuai Wang, Longhao Li, Zhao Guo, Tianlun Zuo, Shuiyuan Wang, Hongfei Xue, Chengyou Wang, Qing Wang, et al. Wenetspeech-chuan: A large-scale sichuanese corpus with rich annotation for dialectal speech processing. *arXiv preprint arXiv:2509.18004*, 2025.

Alex Graves. Sequence Transduction with Recurrent Neural Networks, 2012. URL http://arxiv.org/abs/1211.3711.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition. In *Proc. Speech and Computer*, pp. 267–278, 2020. URL https://doi.org/10.1007/978-3-030-60276-5_27.

Longhao Li, Zhao Guo, Hongjie Chen, Yuhang Dai, Ziyu Zhang, Hongfei Xue, Tianlun Zuo, Chengyou Wang, Shuiyuan Wang, Jie Li, et al. Wenetspeech-yue: A large-scale cantonese speech corpus with multi-dimensional annotation. *arXiv preprint arXiv:2509.03959*, 2025.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech*, pp. 498–502, 2017. URL https://doi.org/10.21437/Interspeech.2017-1386.

Bingshen Mu, Pengcheng Guo, Zhaokai Sun, Shuai Wang, Hexin Liu, Mingchen Shao, Lei Xie, Eng Siong Chng, Longshuai Xiao, Qiangze Feng, and Daliang Wang. Summary on The Multilingual Conversational Speech Language Model Challenge: Datasets, Tasks, Baselines, and Methods. In *Proc. ICASSP*, 2026a.

Bingshen Mu, Xian Shi, Xiong Wang, Hexin Liu, Jin Xu, and Lei Xie. LLM-ForcedAligner: A Non-Autoregressive and Accurate LLM-Based Forced Aligner for Multilingual and Long-Form Speech, 2026b. URL https://arxiv.org/abs/2601.18220.

OpenAI. Hello GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proc. ICML*, pp. 28492–28518, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Elena Rastorgueva, Vitaly Lavrukhin, and Boris Ginsburg. NeMo Forced Aligner and its application to word alignment for subtitle generation. In *Proc. Interspeech*, pp. 5257–5258, 2023. URL https://www.isca-archive.org/interspeech_2023/rastorgueva23_interspeech.html.

Xian Shi, Yanni Chen, Shiliang Zhang, and Zhijie Yan. Achieving timestamp prediction while recognizing with non-autoregressive end-to-end ASR model. In *Proc. NCMMSC*, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni Technical Report, 2025b. URL https://arxiv.org/abs/2509.17765.

Z.ai. Glm asr 2512. https://docs.z.ai/guides/audio/glm-asr-2512, 2025. Accessed: 2026-01-26.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *Proc. ICASSP*, pp. 6182–6186, 2022.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Appendix

Table A.1: **Evaluation on English, Chinese and a range of Chinese dialect benchmarks. As a member of Qwen3-ASR family, Qwen3-ASR-Flash-1208 serves as an API and its results are for reference in the table.**

	Qwen3-ASR-0.6B	**Qwen3-ASR-1.7B**	**Qwen3-ASR-Flash-1208**
English (en)			
LibriSpeech *clean* | *other*	2.11 | 4.55	1.63 | 3.38	1.33 | 2.40
GigaSpeech	8.88	8.45	8.82
CV-en	9.92	7.39	6.06
Fleurs-en	4.39	3.35	2.72
MLS-en	6.00	4.58	3.63
Tedlium	3.85	4.50	4.84
VoxPopuli	9.96	9.15	8.45
Chinese (zh)			
WenetSpeech *net* | *meeting*	5.97 | 6.88	4.97 | 5.88	4.60 | 5.80
AISHELL-2-test	3.15	2.71	2.53
SpeechIO	3.44	2.88	2.62
Fleurs-zh	2.88	2.41	2.38
CV-zh	6.89	5.35	4.45
Chinese Dialect			
KeSpeech	7.08	5.10	3.28
Fleurs-yue	5.79	3.98	3.50
CV-yue	9.50	7.57	4.86
CV-zh-tw	5.59	3.77	3.30
WenetSpeech-Yue *short* | *long*	7.54 | 9.92	5.82 | 8.85	5.84 | 8.20
WenetSpeech-Chuan *easy* | *hard*	13.92 | 24.45	11.99 | 21.63	11.52 | 20.82

Table A.2: **Evaluation of Qwen3-ASR on open-source multilingual benchmarks. As a member of Qwen3-ASR family, Qwen3-ASR-Flash-1208 serves as an API and its results are for reference in the table.**

(a) **MLS, CommonVoice, and MLC-SLM.**

	Qwen3-ASR -0.6B	Qwen3-ASR -1.7B	Qwen3-ASR -Flash-1208
MLS			
da	16.79	11.73	7.58
de	9.52	6.05	4.11
en	6.04	4.58	3.63
es	7.19	4.63	3.29
fr	8.55	5.26	3.16
it	19.21	13.20	7.88
pl	26.09	15.26	9.76
pt	12.16	7.71	6.83
CommonVoice			
ar	45.99	37.97	33.86
de	9.44	5.85	3.53
en	9.92	7.39	6.06
es	7.16	4.65	3.14
fr	12.25	8.56	5.88
it	10.16	5.40	3.21
ja	14.96	11.64	9.31
ko	8.48	5.88	3.82
pt	11.30	7.10	5.42
ru	14.07	8.28	5.73
yue	9.50	7.57	4.86
zh	6.89	5.35	4.45
zh_tw	5.59	3.77	3.30
MLC-SLM			
de	19.78	17.19	15.76
en	7.44	6.41	6.55
es	13.89	11.07	9.31
fr	22.96	20.75	22.98
it	21.31	16.75	14.93
ja	14.74	11.80	9.74
ko	10.31	8.61	8.09
pt	34.97	26.64	28.14
ru	19.24	15.17	13.16
th	19.51	14.34	19.66
vi	17.67	14.92	13.11

(b) **Fleurs.**

	Qwen3-ASR -0.6B	Qwen3-ASR -1.7B	Qwen3-ASR -Flash-1208
Fleurs			
ar	25.51	16.98	14.78
cs	47.67	22.42	18.68
da	36.36	21.00	11.85
de	6.48	3.92	3.03
el	49.67	28.08	13.85
en	4.39	3.35	2.72
es	4.94	3.36	2.68
fa	53.76	29.90	18.37
fi	46.59	25.23	12.21
fil	36.10	24.29	19.17
fr	7.72	4.75	3.44
hi	19.12	17.15	13.77
hu	59.47	34.22	21.77
id	7.92	5.16	3.65
it	4.99	2.41	1.60
ja	8.33	5.20	3.09
ko	3.72	2.57	2.07
mk	37.26	19.05	–
ms	17.66	10.39	11.37
nl	14.02	7.04	4.35
pl	24.71	12.54	7.24
pt	6.21	3.92	3.18
ro	44.26	20.70	10.45
ru	9.91	5.99	4.81
sv	35.87	19.36	15.02
th	8.34	6.32	5.53
tr	16.18	9.47	6.13
vi	8.52	5.55	3.64
yue	5.79	3.98	3.50
zh	2.88	2.41	2.38