PRONCOACH: PERCEIVE, REASON, AND CORRECT PRONUNCIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pronunciation correction aims to improve speech clarity through realistic pronunciation recognition and targeted pronunciation guidance. Reliable pronunciation coaching faces three challenges: (1) languages represent pronunciation differently; (2) fine-grained pronunciation-correction data and phone-level annotations are scarce, while ambiguity in phone perception leads to unstable expert annotations; and (3) feedback must account for weak forms, tone changes, accent variation, and other context-dependent phenomena. We present PRONCOACH, a pronunciation coaching framework for Mandarin and English audio without reference transcripts. We decompose coaching into a structured chain-of-thought process covering intended-content recovery, realized-pronunciation inference, contextual comparison, per-phone correction, intervention decisions, and feedback generation. Based on this decomposition, we fine-tune language-specific phonetic-perception models on expert annotations to infer probabilistic candidate realizations from acoustic and canonical-pronunciation evidence. An audio-language model (ALM) then uses the recording, intended content, and candidate evidence to generate aligned diagnoses and coaching feedback. These annotations supervise both complete coaching responses and intermediate tasks. To address the lack of a dedicated pronunciation-correction benchmark, we construct PRONCOACH-BENCH with two complementary tracks. The Diagnostic Track evaluates correction localization, heard-realization agreement, and false corrections, jointly testing phonetic perception and correction decisions; the Clean-Speech Track tests correction restraint by measuring whether models avoid unnecessary feedback. The main PRONCOACH model achieves a heard-based overall F1 of 0.7528, compared with 0.4139 for the strongest evaluated public ALM baseline. Supplementary English evaluations yield the lowest phoneme error rate among the compared systems on two of three corpora. Furthermore, PRONCOACH demonstrates emergent cross-lingual correction capabilities, enabling it to perform pronunciation rectification across Mandarin and English representations. PRONCOACH-BENCH will be publicly released soon.

1 INTRODUCTION

Pronunciation correction aims to improve speech clarity by identifying how a learner spoke, locating pronunciation problems, and explaining how to correct them (Neri et al., 2002; Witt & Young, 2000). It covers accent-related deviations and other pronunciation issues, while recognizing that accentedness, comprehensibility, and intelligibility are distinct (Munro & Derwing, 1995; Levis, 2005). Pronunciation coaching turns these diagnoses into actionable explanations and practice (Lee et al., 2015), supporting everyday learning and preparation for speaking examinations.

Reliable coaching faces three challenges. First, languages use different phonetic and teaching conventions: English commonly uses IPA, whereas Mandarin commonly uses pinyin-based onset, rime, and tone annotations. Second, annotations of realized pronunciation, error locations, and target pronunciation are scarce, and ambiguity can make experts disagree on the same recording (Guo et al., 2026). Third, stress, weak forms, tone sandhi, and accent variation require contextual judgment (Bell et al., 2009; Zhang et al., 2015; Munro & Derwing, 1995). A dictionary mismatch may be acceptable, while correct word recognition may conceal a pronunciation problem. Correction must therefore account for both pronunciation uncertainty and context.

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

Existing tasks do not capture these requirements jointly. Pronunciation assessment and mispronunciation detection and diagnosis produce scores or error labels, often requires given intended text or canonical pronunciation (Witt & Young, 2000; Vidal et al., 2023; Ryu et al., 2023; Wang et al., 2024). Audio-language models (ALMs) have been trained to predict realized pronunciation from audio and supplied canonical pronunciation (Wu et al., 2024; 2025); feedback generators turn supplied diagnoses into articulatory explanations (Zhong et al., 2024). These outputs alone do not solve pronunciation correction: the model must recover intended content, preserve plausible pronunciations under ambiguity, distinguish errors from acceptable variants, and explain warranted corrections. A transcript can omit the phonetic differences needed for these decisions. Figure 1 illustrates this task-level distinction for pronunciation correction.

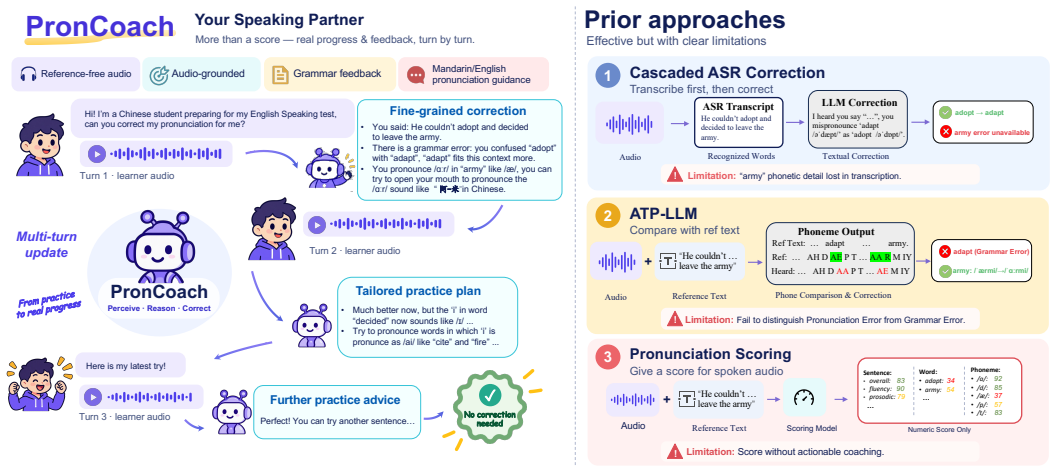


Figure 1: Pronunciation correction links intended content and realized pronunciation to warranted corrections and actionable guidance. The comparison illustrates why transcription, scores, or phone predictions alone do not fulfill this task.

To our knowledge, we provide the first formulation and dedicated solution of this joint pronunciation-correction task for Mandarin and English audio without reference transcripts. Our framework, PRONCOACH, connects intended-content recovery and realized-pronunciation perception to diagnosis, targeted correction, and practice guidance. It also provides grammatical feedback when pronunciation deviations affect expression and guides practice for recurring problems across readings. We observe bidirectional cross-language pronunciation rendering: English through Mandarin forms and Mandarin through English forms. The framework is designed to preserve acoustic pronunciation evidence for specific, context-appropriate feedback.

We organize supervision around structured reasoning. *Intended-content recovery* establishes the learner's intended words; their expected pronunciation is the *canonical pronunciation*. *Realized-pronunciation inference* estimates how those words were spoken. Because phonetic categories span overlapping acoustic ranges, even experts may disagree; pronunciation probabilities retain plausible realizations instead of forcing one interpretation. Language-specific phonetic-perception models trained on expert annotations extend this evidence to a larger speech pool. *Contextual comparison* aligns canonical and realized pronunciations with text positions while accounting for acceptable contextual forms. *Per-phone correction* proposes adjustments; *Intervention decisions* select those warranting feedback; *Feedback generation* produces explanations and practice. A teacher ALM checks the audio, intended content, and pronunciation evidence to construct aligned supervision for these intermediate tasks and complete coaching responses.

PRONCOACH-BENCH evaluates this task beyond phone recognition. Its Diagnostic Track jointly measures correction localization, realized-pronunciation agreement, and false corrections; its Clean-Speech Track tests restraint on correctly pronounced speech. The main PRONCOACH model achieves overall F1 of 0.7528 versus the strongest evaluated public ALM's 0.4139, a 0.3389 improvement across diagnosis and clean-speech behavior. Phones extracted from complete coaching responses yield the lowest phoneme error rate (PER) on SpeechOcean762 and EpaDB among the

compared systems. These evaluations separately test pronunciation descriptions and correction decisions. The benchmark and evaluation data will be released.

2 Related Work

Realized-Pronunciation Modeling and Mispronunciation Diagnosis. Pronunciation assessment and mispronunciation detection and diagnosis (MDD) use likelihood-based scoring, self-supervised representations, multitask learning, and pitch modeling for phone-level scores and diagnoses (Witt & Young, 2000; Vidal et al., 2023; Ryu et al., 2023; Wang et al., 2024). MixGoP models pronunciation variation through multiple acoustic subclusters per phone (Choi et al., 2025); ATP-LLM predicts realized pronunciations from audio and canonical pronunciations, and PP-ATP adds potential pronunciations (Wu et al., 2024; 2025). HuPER combines acoustic-phonetic evidence with linguistic knowledge for adaptive perception (Guo et al., 2026). For the version used here, we address limited English coverage and the lack of direct Mandarin annotation support through expert-supervised language-specific fine-tuning; Section 3.2 details our treatment of annotation uncertainty.

Cross-lingual wav2vec2 phoneme recognition (Xu et al., 2021) and ZIPA (Zhu et al., 2025) provide phone-sequence predictions; ZIPA's CR and CR-NS variants use consistency-regularized CTC and further noisy-student training, respectively. MPL-MDD uses momentum pseudo-labeling to exploit unlabeled speech for mispronunciation detection (Yang et al., 2022), while CROTTC-IF separates acoustic recognition from explicit canonical prompting (Geng et al., 2026). MDD-LLM (Llama-3.2)[1] provides a language-model-based system for predicting perceived phone sequences from speech. Their outputs address phonetic recognition or diagnosis; the task defined here additionally requires context-appropriate correction and guidance.

Audio-Language Models for Pronunciation Coaching. General-purpose audio-language models support speech-conditioned instruction following. The Qwen-Omni models combine audio perception with language generation, producing text or speech responses to multimodal inputs (Xu et al., 2025a;b; Team, 2026). These capabilities provide the foundation for our domain model. Current ALMs, including Gemini (Team et al., 2023), Doubao Seed Lite (Seed, 2026), and Muse Spark[2], can also provide correction feedback. Zhong et al. generate feedback from detected phones and articulatory features and evaluate its comprehensibility and helpfulness (Zhong et al., 2024).

Pronunciation-Coaching Supervision and Evaluation. Pronunciation-learning supervision and evaluation emphasize scoring and phone-level diagnosis, with limited coverage of complete coaching responses. SpeechOcean762 offers sentence-, word-, and phone-level expert assessments (Zhang et al., 2021); L2-ARCTIC contains English speech from multiple first-language backgrounds with manual phone annotations (Zhao et al., 2018); EpaDB provides fine-grained realized-pronunciation annotations (Vidal et al., 2019). Our benchmark additionally tests localization, faithful realization descriptions, and unnecessary intervention (Section 5).

3 PronCoach

3.1 Task Overview

PronCoach takes learner audio and optional practice instructions or dialogue context, and returns diagnoses, explanations, and practice guidance without a supplied reference transcript. Figure 2 groups supervision construction into five stages: intended-content recovery (green), canonical-pronunciation inference (yellow), realized-pronunciation inference (blue), contextual comparison and per-phone correction (red), and intervention decisions and feedback generation (purple). Thinking responses expose the intermediate comparisons; Instruct responses present the resulting diagnoses and guidance directly.

[1] https://huggingface.co/Haopeng/MDD-LLM-Llama3.2-1B-L2-ARCTIC

[2] https://research.meta.ai/blog/multimodal-intelligence-of-muse-spark-1-2

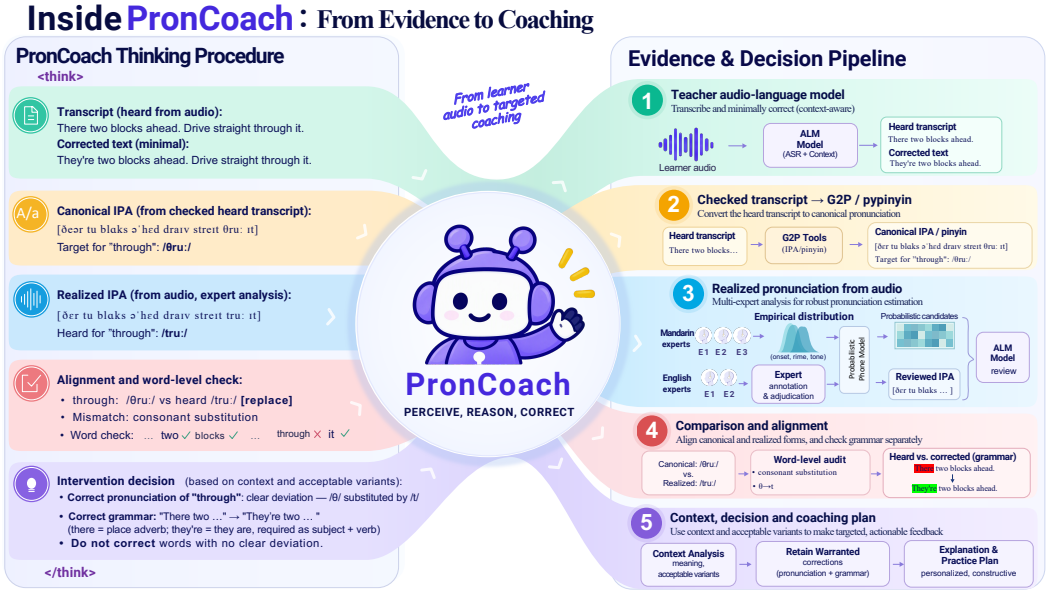


Figure 2: Supervision construction for PRONCOACH: the five-stage reasoning (left) and annotation process (right) align pronunciation probabilities, contextual diagnoses, and coaching feedback.

3.2 CONTENT AND PRONUNCIATION PERCEPTION

Intended-content recovery addresses accent-related transcription errors that change lexical meaning. Realizing /v/ as /w/ in "He wore a vest over his shirt", for example, may cause *vest* to be transcribed as *west*. The clothing context supports recovering *vest*, while /w/ remains realized pronunciation evidence. During annotation, Gemini 3.1 Pro Preview checks the recording and available transcript in context to recover the intended lexical content.

Language-specific grapheme-to-phoneme (G2P) processing converts the recovered text into canonical pronunciation, preserving text-unit associations. English uses espeak-ng[3] through phonemizer[4] to produce IPA with lexical stress; primary/secondary stress marks become independent tokens, and Unicode (NFC) normalization precedes phoneme segmentation. Mandarin uses pypinyin[5] for tone-bearing pinyin with onset, rime, and tone components. A grammar-corrected version is stored separately for grammar diagnosis.

Multiple experts annotate realized pronunciation. For Mandarin, three experts annotate independently; English uses two annotators and a third expert for adjudication. Pronunciation probabilities preserve these alternatives for review: Mandarin retains independent judgments as empirical onset, rime, and tone candidate frequencies; English disagreements are adjudicated into reviewed IPA sequences. We fine-tune language-specific HuPER (Guo et al., 2026) models on expert annotations to extend pronunciation coverage and estimate candidate probabilities. Both combine canonical pronunciation with acoustic evidence from WavLM for Mandarin (Chen et al., 2022) and HuBERT for English (Hsu et al., 2021). Mandarin uses cross-entropy against expert pronunciation probabilities; English learns keep, delete, substitute, and insert operations from aligned canonical and realized pronunciations.

The fine-tuned models perform *realized-pronunciation inference* on a larger transcribed speech pool, followed by ALM review against the recording. Probabilities guide review without overriding acoustic evidence, which may support a lower-probability candidate. Reviewed text and pronunciations feed diagnosis; expert annotations, model estimates, and ALM-reviewed records remain separate from each other and from rendered training targets (Section 3.4).

[3]https://github.com/espeak-ng/espeak-ng
[4]https://github.com/bootphon/phonemizer
[5]https://github.com/mozillazg/python-pinyin

3.3 Pronunciation and Grammar Diagnosis

Pronunciation diagnosis compares canonical and realized pronunciations to identify candidate problems; a difference alone need not warrant correction. Grammar diagnosis separately records changes required by word or sentence-structure errors, including when the words are pronounced correctly.

Grammar diagnosis records differences between the recovered transcript and its minimally grammar-corrected version. If revision changes *There* to *They're*, it records a grammatical issue; pronunciation analysis still compares the heard *There* with its own canonical pronunciation, avoiding spurious phonetic insertions, deletions, or substitutions introduced by grammar revision.

Contextual comparison links each recovered text unit to its canonical pronunciation and reviewed realized pronunciation. Mandarin compares character-level onset, rime, and tone. For English, the construction program aligns phonetic tokens, records substitutions, insertions, and deletions with positions in both sequences, and removes equivalent-notation differences. Stress marks are excluded from alignment; the ALM checks lexical stress word by word against the audio.

Per-phone correction combines these comparisons with ALM diagnoses into candidate correction records. A pass over the entire recovered transcript links diagnoses to Mandarin characters or English words. Flagged English words retain target and heard pronunciations and issue type; others receive "no clear deviation" without additional word-level IPA. Unassignable diagnoses are stored separately and retained in training feedback. In long utterances, shared text positions help identify skipped or misplaced corrections and keep the comparison and feedback aligned.

The intervention criteria (Section 3.4) select contrasts for feedback. The comparison trace and final feedback share the retained position-linked records; positional consistency checks exclude unlocalized diagnoses, which remain in feedback without counting as successfully localized corrections.

3.4 Correction Decisions and Coaching Feedback

The ALM judges corrections from the recording, heard and corrected texts, and pronunciation evidence in word-level context. English prompts follow IELTS and TOEFL pronunciation criteria while allowing specified function-word variants and natural reductions; Mandarin accounts for tone sandhi, neutral tones, erhua, and connected-speech changes. These criteria select differences for feedback; sequence mismatch alone is insufficient.

Each retained pronunciation correction specifies the text unit, realized pronunciation, canonical pronunciation, and issue type; grammar corrections separately record expression changes. From these diagnoses, the ALM generates an overall comment, concrete adjustments, and a practice plan with articulatory guidance, sound or word contrasts, and further readings. When intervention is unwarranted, correction records remain empty.

The construction program checks agreement in text positions, pronunciations, and correction entries between the comparison trace and final response. Rendered training sequences retain diagnosed realizations at correction positions and normalize other positions to primary canonical forms; this output convention does not overwrite the original pronunciation evidence. We standardize pinyin, IPA, and word boundaries, verify sequence–text correspondence, and remove equivalent-notation differences and rule-excluded entries. Pronunciation accuracy concerns the sounds reported in the response; correction measures concern the selected text units and whether feedback is warranted.

Tongue twisters, confusable-sound readings, and CosyVoice3 (Du et al., 2025) synthesized contrasts supplement the speech pool and undergo the same annotation procedure. Multi-turn examples pair each separately analyzed new recording and diagnosis with preceding feedback, continuing the practice in Figure 1. These records provide supervision; inference requires no user-supplied transcript, HuPER prediction, or expert annotation. Appendix A details implementation and annotation.

4 PronCoach Training

We train PronCoach on the constructed data using supervised fine-tuning (SFT), followed by group relative policy optimization (GRPO).

Multitask supervised fine-tuning. We initialize PRONCOACH from Qwen3.5-Omni and train on eight task types from data construction: transcription, accent description, grammar checking, canonical pronunciation prediction, realized pronunciation prediction, pronunciation-error detection, structured diagnosis, and complete coaching feedback. SFT learns these constructed targets through autoregressive token prediction.

Approximately 20k training instances comprise 13k single-turn and 7k multi-turn examples, with equal numbers of Mandarin and English correction tasks. We render the same pronunciation and correction annotations in two modes: *Thinking* includes explicit pronunciation comparisons and correction reasoning before feedback; *Instruct* directly presents diagnoses, comments, and practice plans. Both preserve the same correction targets. Training mixes Thinking inline responses, without a separate reasoning block, and Instruct responses 1:1.

Relative to *Main*, *Discrete targets* replaces probabilistic realized-pronunciation candidates in the ALM's annotation input with a single realization, changing supervision evidence rather than only response format. *w/o Thinking data* omits explicit intermediate comparisons from training while retaining final feedback; data, initialization, sample count, training steps, and decoding settings otherwise match Main. Main retains Thinking supervision even when evaluated in Instruct mode. *Scale* expands the balanced data, preserving Main's Mandarin–English balance and proportions across tasks and response forms. Approximately 370k examples with 120k unique recordings are included in *Scale* model training. Appendix B.3 details these comparisons.

Reinforcement learning for pronunciation diagnosis. After SFT, we apply group relative policy optimization (GRPO) (Shao et al., 2024) to constrain output format and reduce missed corrections in Mandarin and English diagnosis. Rewards computed from the final structured answer combine recall-oriented diagnosis, JSON validity, and canonical-pronunciation consistency:

$$\mathcal{R} = 0.85F_{1.2} + 0.10I_{\text{parse}} + 0.05I_{\text{canonical}}. \tag{1}$$

The diagnosis term $F_{1.2}$ combines precision P_{tr} (support for reports) and recall R_{tr} (coverage of human-reference corrections), crediting matched positions and correctly described canonical and realized pronunciations.

$$F_\beta = \frac{(1+\beta^2)P_{\text{tr}}R_{\text{tr}}}{\beta^2 P_{\text{tr}} + R_{\text{tr}}}, \qquad \beta = 1.2.$$

The recall emphasis is intended to reduce missed corrections, with precision retained to discourage unsupported reports. Appendix B.5 specifies the matching credits and reference weights.

The format term I_{parse} is 1 for a parsed JSON object and 0 otherwise, supporting automatic extraction and scoring; an unparseable answer receives zero total reward. $I_{\text{canonical}}$ is 1 only when the complete canonical pronunciation sequence matches the normalized reference, and 0 otherwise. This utterance-wide consistency check complements position-level diagnosis matching.

GRPO samples multiple responses per input and updates the policy using group-normalized reward advantages and KL regularization against a reference policy. It optimizes structured diagnosis; SFT supervises explanations and practice guidance.

5 PRONCOACH-BENCH

PRONCOACH-BENCH contains 688 recordings in complementary Diagnostic and Clean-Speech Tracks, jointly testing pronunciation-problem identification, accurate realized-pronunciation descriptions, and restraint when no correction is needed.

Diagnostic contains 510 expert phonetically annotated recordings: 296 Mandarin and 214 English utterances. Mandarin annotations retain onset, rime, and tone candidates; English provides reviewed IPA. Gemini 3.1 Pro Preview combines each recording, transcript, and human pronunciation annotations into reference corrections specifying the text unit, target pronunciation, realized pronunciation, and issue type. We retain only recordings with complete, localizable reference records. Structured answers are parsed directly; GPT-4o-mini converts free-form answers into the same format. Shared normalization and filtering precede alignment with reference positions.

We score each language separately. Recall R is the fraction of reference correction positions found; localization precision P_{loc} is the fraction of localizable predictions matching these positions. Their harmonic mean, $F1_{\mathrm{loc}}$, assesses text-unit selection, accounting for missed and incorrect corrections. At correctly localized positions, $\mathrm{ACC}_{\mathrm{hear}}$ measures predicted–reference realized-pronunciation agreement: a shared normalized pronunciation candidate in Mandarin, or matching normalized whole-word IPA in English. We combine reference coverage and realized-pronunciation agreement as

$$F1^{\mathrm{hear}} = \frac{2R\,\mathrm{ACC}_{\mathrm{hear}}}{R + \mathrm{ACC}_{\mathrm{hear}}}.$$

FC divides the number of correction positions absent from the reference, plus complete corrections unlocalizable in the transcript, by the number of evaluated positions without reference corrections. It measures unsupported reports, including unlocalizable reports excluded from localization precision.

Clean-Speech contains 178 recordings designated as standard pronunciation: 81 standard broadcasting Mandarin utterances and 97 English utterances from IELTS and TOEFL materials. Each receives a pronunciation-correction request in its language, testing recognition that no correction is needed. For each language, ACC_c is the proportion receiving valid responses without pronunciation or grammar corrections. Missing or unparseable responses remain in the denominator as non-clean. For aggregation, $\overline{\mathrm{ACC}}_c$ averages both languages' accuracies, weighted by their recording counts.

For the overall score, we obtain $F1_D^{\mathrm{hear}}$ by averaging Mandarin and English $F1^{\mathrm{hear}}$, weighted by retained reference correction counts, then combine diagnostic and Clean-Speech performance:

$$F1_{\mathrm{all}}^{\mathrm{hear}} = \frac{n_D F1_D^{\mathrm{hear}} + n_C \dfrac{2\overline{\mathrm{ACC}}_c}{1 + \overline{\mathrm{ACC}}_c}}{n_D + n_C},$$

where $n_D = 510$ and $n_C = 178$ count Diagnostic and Clean-Speech recordings. The score summarizes reference coverage, realized-pronunciation agreement, and clean-speech performance; $F1_{\mathrm{loc}}$ and FC separately assess correction-selection errors. English-only comparisons use English track scores and recording counts. Complete definitions, weights, and failure rules appear in Appendix C.

6 Experiments

We evaluate PronCoach on pronunciation correction against ALM baselines on PronCoach-Bench, phonetic baselines on three public English corpora for phone recognition, and the same phonetic baselines on PronCoach-Bench for correction. Ablations examine pronunciation probabilities in annotation, Thinking supervision, and training-data scale.

6.1 Evaluation setup

We evaluate PronCoach with Instruct responses against five public ALMs: Doubao Seed 2.0 Lite, Gemini 3.8 Flash, Gemini 3.1 Pro Preview, Muse Spark 1.2, and Qwen3.5-Omni-plus (Table 1). Each receives a recording and a pronunciation-correction request without a reference transcript. Recordings, reference corrections, filters, and scoring are shared; request wording and response formats differ. Appendix C.2 details model inputs and response handling.

To assess pronunciation perception beyond our own benchmark, we compare PronCoach with phonetic baselines on test set of L2-ARCTIC (Zhao et al., 2018), SpeechOcean762 (Zhang et al., 2021), and EpaDB (Vidal et al., 2019) using phoneme error rate (PER). We also compare their pronunciation-correction performance on the English subsets of PronCoach-Bench (Table 2). Evaluation details are provided in Appendices C.6 and C.3.

Table 3 compares the training variants defined in Section 4. All use Instruct responses at evaluation and share references, filters, localization, and metrics; Appendix B.3 gives details.

6.2 Overall performance

PronCoach outperforms the five public ALM baselines in overall F1 (Table 1), reaching 0.7528 compared with 0.4139 for the strongest baseline. It also achieves higher localization F1 and realized-

Table 1: Reference-free ALMs on PRONCOACH-BENCH: per-language Diagnostic and Clean-Speech metrics, with Overall spanning both languages and tracks. Appendix D gives full results. All metrics are higher the better except FC.

Model	Language	Diagnostic			Clean-Speech	Overall
		$\mathrm{ACC}_{\mathrm{hear}} \uparrow$	$F1_{\mathrm{loc}} \uparrow$	FC $\downarrow$	$\mathrm{ACC}_c \uparrow$	$F1^{\mathrm{hear}}_{\mathrm{all}} \uparrow$
Doubao Seed 2.0 Lite	Mandarin	0.5274	0.1941	**0.3331**	0.0370	0.3073
	English	<u>0.1974</u>	<u>0.2222</u>	0.3672	0.0206	
Gemini 3.8 Flash	Mandarin	0.5370	0.1430	0.5570	0.0247	0.3777
	English	0.1489	0.2108	0.4869	<u>0.5979</u>	
Gemini 3.1 Pro Preview	Mandarin	0.3158	0.1319	0.6850	0.0247	0.3941
	English	0.1860	0.2161	<u>0.1421</u>	**0.9072**	
Muse Spark 1.2	Mandarin	0.3846	0.1224	0.6568	0.0741	0.2936
	English	0.0485	0.1737	0.6852	0.4227	
Qwen3.5-Omni-plus	Mandarin	<u>0.5497</u>	<u>0.2130</u>	<u>0.3440</u>	**0.3457**	<u>0.4139</u>
	English	0.1000	0.1816	0.5337	0.2784	
PRONCOACH	Mandarin	**0.9799**	**0.3124**	0.3501	<u>0.1728</u>	**0.7528**
	English	**0.7981**	**0.5762**	**0.0605**	0.5567	

Table 2: Comparison with open-source models on English phoneme recognition and correction ability. PER (%; lower is better) uses separate corpora and is excluded from Overall; SO762 denotes SpeechOcean762. Input conditions and the Diagnostic/Clean-Speech scope follow Section 6.1.

Model	L2-ARCTIC	SO762	EpaDB	PronCoach Bench				
	Phone Recognition			Diagnostic			Clean-Speech	Overall
	PER (%) $\downarrow$			$\mathrm{ACC}_{\mathrm{hear}} \uparrow$	$F1_{\mathrm{loc}} \uparrow$	FC $\downarrow$	$\mathrm{ACC}_c \uparrow$	$F1^{\mathrm{hear}}_{\mathrm{all}} \uparrow$
W2V2-eSpeak	23.34	35.80	<u>34.77</u>	0.1042	<u>0.2609</u>	0.4906	0.0103	0.1335
ZIPA-CR-Large	24.26	42.07	38.71	0.0789	0.2375	0.4196	0.2371	0.2164
ZIPA-CR-NS-Large	23.75	44.30	40.90	0.0893	0.2383	<u>0.4084</u>	<u>0.2784</u>	0.2436
CROTTC-IF	34.42	94.60	68.61	0.0959	0.2135	0.6540	0.0000	0.1183
MDD-LLM	<u>12.67</u>	<u>34.89</u>	37.86	0.2239	0.2550	0.4638	0.0000	0.2381
MPL-MDD	**11.30**	35.61	39.16	<u>0.3139</u>	0.2453	0.5031	0.0000	<u>0.3078</u>
PRONCOACH	13.86	**15.76**	**25.59**	**0.7981**	**0.5762**	**0.0605**	**0.5567**	**0.6903**

pronunciation agreement in both languages. The gains extend from identifying the affected text units to accurately describing their realized pronunciations at correctly localized positions.

The results are less uniform for unnecessary corrections. In Mandarin, PRONCOACH has slightly higher FC than Doubao and Qwen, and lower Clean-Speech accuracy than Qwen. In English, it achieves the lowest Diagnostic FC, while Gemini Flash and Pro attain higher Clean-Speech accuracy. In our qualitative use, both Gemini models tended to withhold corrections, limiting the usefulness of their coaching feedback. Their lower Diagnostic localization F1 and realized-pronunciation agreement further show that avoiding corrections on clean speech does not by itself establish strong coaching performance. We therefore consider both tracks jointly: a pronunciation coach should identify and explain problems that require correction while avoiding unnecessary corrections to acceptable pronunciations.

PRONCOACH achieves the lowest PER on SpeechOcean762 and EpaDB, while MDD-LLM and MPL-MDD score lower on L2-ARCTIC (Table 2). Together with the correction results, this demonstrates phonetic-perception and pronunciation-correction performance across the evaluated settings.

Table 3: Training variants on PRONCOACH-BENCH: per-language Diagnostic and Clean-Speech metrics and bilingual Overall. Main is the reference; Section 4 and Appendix B.3 specify the variants.

Model	Setting	Language	Diagnostic Track			Clean-Speech	Overall
			$\mathrm{ACC}_{\mathrm{hear}}\uparrow$	$F1_{\mathrm{loc}}\uparrow$	$\mathrm{FC}\downarrow$	$\mathrm{ACC}_{c}\uparrow$	$F1_{\mathrm{all}}^{\mathrm{hear}}\uparrow$
PRONCOACH	Discrete targets	Mandarin	0.9549	0.6348	0.0234	0.5309	0.6380
		English	0.5412	0.4582	0.0779	0.4330	
PRONCOACH	w/o Thinking data	Mandarin	0.7421	0.1805	0.6105	0.0000	0.4240
		English	0.1928	0.3333	0.1833	0.0000	
PRONCOACH	Scale	Mandarin	0.9876	0.3932	0.2716	0.3333	0.7837
		English	0.8897	0.5620	0.1453	0.2165	
PRONCOACH	Main	Mandarin	0.9799	0.3124	0.3501	0.1728	0.7528
		English	0.7981	0.5762	0.0605	0.5567	

6.3 TRAINING VARIANTS AND ABLATION STUDY

Pronunciation probabilities. Compared with Discrete targets, Main achieves higher realized-pronunciation agreement in both languages, improves all reported English Diagnostic and Clean-Speech metrics, and raises overall F1 from 0.6380 to 0.7528. Discrete targets achieves higher Mandarin localization F1, lower FC, and higher Clean-Speech accuracy. As discussed in Section 3.2, phonetic symbols cover broad acoustic ranges, so ambiguous pronunciations can receive different expert judgments. Discrete targets retains only the highest-probability realization, discarding alternative hypotheses before the ALM reviews the audio. Pronunciation probabilities allow the ALM to select a lower-probability realization when it is supported by the recording and to construct more informative pronunciation supervision. This supports realized-pronunciation inference, while contextual comparison remains necessary to determine which differences warrant correction.

Removing Thinking data. Removing Thinking data lowers overall F1 from 0.7528 to 0.4240, weakens pronunciation diagnosis, and reduces Clean-Speech accuracy to zero in both languages. Thinking examples explicitly show how pronunciation comparisons lead to correction decisions, helping the model learn which differences require feedback. Both configurations are evaluated in Instruct mode. These results support the effectiveness of our Thinking data decomposition as training supervision.

Training-data scale. Scale improves all reported Mandarin metrics and raises overall F1 from 0.7528 to 0.7837. In English, realized-pronunciation agreement improves, but localization F1 and Clean-Speech accuracy decline, while FC increases. These results suggest that additional training data strengthen pronunciation perception, although the benefits differ across languages.

7 CONCLUSION

We present PRONCOACH, an ALM-based framework for Mandarin and English pronunciation correction and guidance. Language-specific phonetic models trained on expert annotations provide pronunciation probabilities, preserving plausible realizations for audio-based review. Our Thinking decomposition connects this evidence to contextual correction decisions and supports the generation of targeted Instruct feedback. We introduce PRONCOACH-BENCH, whose Diagnostic and Clean-Speech Tracks jointly evaluate pronunciation diagnosis and unnecessary corrections. Experiments show stronger overall correction performance than the evaluated baselines and lower PER on most evaluated open-source English datasets. We also observe cross-language pronunciation rendering between Mandarin and English. Future work will improve phone and tone diagnosis, reduce unnecessary corrections, and assess learner progress through continued practice.

ETHICS STATEMENT

Pronunciation differences should not be used to judge a speaker's identity or ability. Coaching targets issues relevant to the learning goal while respecting context-appropriate variation. Voice recordings may identify speakers; their use and redistribution, including derived annotations, remain subject to source licenses and consent. Feedback should allow disagreement and correction rather than present uncertain judgments as absolute standards.

REPRODUCIBILITY STATEMENT

The paper details pronunciation supervision, example construction, training objectives, and evaluation; the appendix specifies metrics and auxiliary HuPER sampling and reference rules. Each language retains its human annotation format, with human references separate from recognizer-generated estimates. Reproduction must specify the model, data split, prompt, and decoding settings together, as each can affect corrections.

AI USE STATEMENT

Generative AI assisted research discussions, code and data analysis, literature search, drafting, and language revision. ALM-generated corrective text is part of the studied data construction and remains distinct from human phone annotations. The authors are responsible for verifying claims, results, and references and for the final manuscript.

REFERENCES

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111, 2009.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/jstsp.2022.3188113.

Kwanghee Choi, Eunjung Yeo, Kalvin Chang, Shinji Watanabe, and David R. Mortensen. Leveraging allophony in self-supervised speech models for atypical pronunciation assessment. In *Proceedings of NAACL 2025: Human Language Technologies, Volume 1*, pp. 2613–2628, 2025. doi: 10.18653/v1/2025.naacl-long.132.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Haopeng Geng, Longfei Yang, Xi Chen, Haitong Sun, Daisuke Saito, and Nobuaki Minematsu. Beyond acoustic sparsity and linguistic bias: A prompt-free paradigm for mispronunciation detection and diagnosis. *arXiv preprint arXiv:2604.22133*, 2026.

Chenxu Guo, Jiachen Lian, Yisi Liu, Baihe Huang, Shriyaa Narayanan, Cheol Jun Cho, and Gopala Anumanchipalli. Huper: A human-inspired framework for phonetic perception, 2026. URL https://arxiv.org/abs/2602.01634.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/taslp.2021.3122291.

Junkyu Lee, Juhyun Jang, and Luke Plonsky. The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, 36(3):345–366, 2015.

John M Levis. Changing contexts and shifting paradigms in pronunciation teaching. *TESOL quarterly*, 39(3):369–377, 2005.

Murray J Munro and Tracey M Derwing. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, 45(1):73–97, 1995.

Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. The pedagogy-technology interface in computer assisted pronunciation training. *Computer assisted language learning*, 15(5):441–467, 2002.

Hyungshin Ryu, Sunhee Kim, and Minhwa Chung. A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning. In *Interspeech 2023*, pp. 959–963, 2023. doi: 10.21437/Interspeech.2023-337.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Jazmín Vidal, Luciana Ferrer, and Leonardo Brambilla. EpaDB: A database for development of pronunciation assessment systems. In *Interspeech 2019*, pp. 589–593, 2019. doi: 10.21437/Interspeech.2019-1839.

Jazmín Vidal, Pablo Riera, and Luciana Ferrer. Mispronunciation detection using self-supervised speech representations. In *9th Workshop on Speech and Language Technology in Education*, pp. 71–75, 2023. doi: 10.21437/SLaTE.2023-15.

Xintong Wang, Mingqian Shi, and Ye Wang. Pitch-aware RNN-T for mandarin chinese mispronunciation detection and diagnosis. In *Interspeech 2024*, pp. 292–296, 2024. doi: 10.21437/Interspeech.2024-2297.

S. M. Witt and S. J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2–3):95–108, 2000. doi: 10.1016/S0167-6393(99)00044-8.

Minglin Wu, Jing Xu, Xixin Wu, and Helen Meng. Prompting large language models with mispronunciation detection and diagnosis abilities. In *Interspeech 2024*, pp. 2990–2994, 2024. doi: 10.21437/Interspeech.2024-1772.

Minglin Wu, Jing Xu, Xueyuan Chen, and Helen Meng. Integrating potential pronunciations for enhanced mispronunciation detection and diagnosis ability in LLMs. In *ICASSP 2025*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887601.

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 preprint arXiv:2503.20215*, 2025a. URL https://arxiv.org/abs/2503.20215.

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. *arXiv preprint arXiv:2509.17765*, 2025b. URL https://arxiv.org/abs/2509.17765.

Qiantong Xu, Alexei Baevski, and Michael Auli. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*, 2021.

Mu Yang, Kevin Hirschi, Stephen D Looney, Okim Kang, and John HL Hansen. Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment. *arXiv preprint arXiv:2203.15937*, 2022.

Caicai Zhang, Quansheng Xia, and Gang Peng. Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an erp study. *Journal of Neurolinguistics*, 33:149–162, 2015.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Interspeech 2021*, pp. 3710–3714, 2021. doi: 10.21437/Interspeech.2021-1259.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-ARCTIC: A non-native english speech corpus. In *Interspeech 2018*, pp. 2783–2787, 2018. doi: 10.21437/Interspeech.2018-1110.

Huihang Zhong, Yanlu Xie, and ZiJin Yao. Leveraging large language models to refine automatic feedback generation at articulatory level in computer aided pronunciation training. In *Interspeech 2024*, pp. 2600–2604, 2024. doi: 10.21437/Interspeech.2024-1005.

Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R Mortensen. Zipa: A family of efficient models for multilingual phone recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19568–19585, 2025.

A Data Construction and Annotation

A.1 Annotation Sources and Pronunciation Evidence

Expert annotations supervise the phonetic-perception models, whose predictions extend pronunciation evidence to additional recordings. The teacher ALM then constructs diagnoses and coaching targets from the recording and its evidence (Table 4). These sources remain separate: teacher-generated corrections do not replace expert hearings or model estimates. Three experts independently annotate each Mandarin recording. For every character, onset, rime, and tone candidates receive empirical frequencies equal to their expert counts divided by three. These component distributions are retained separately. English annotations instead use two experts and an additional adjudicator to resolve disagreements, yielding reviewed IPA sequences.

Table 4: Pronunciation evidence and its role in data construction. Model probabilities describe the model's candidate realizations; they are not additional expert votes.

Source	**Evidence (ZH)**	**Evidence (EN)**	**Use**
Expert annotation	Per-component expert counts	Adjudicated IPA	Phonetic supervision; ALM input
Phonetic-perception model	Probabilistic onset, rime & tone	Candidate IPA	Annotation of additional audio
ALM annotation	Reviewed pronunciations; indexed diagnoses	Word-aligned IPA; diagnoses	Feedback & task targets

The resulting records support transcription, pronunciation perception, diagnosis, and coaching. Tongue twisters, confusable-sound readings, and synthesized contrasts supplement the speech pool and undergo the same annotation procedure. Multi-turn records pair each new recording and diagnosis with preceding feedback; each recording is analyzed anew. A recording may support several task targets, whose coverage and response formats are described in Appendix B. The examples below trace two actual recordings from pronunciation evidence to feedback.

A.2 TEACHER ANNOTATION POLICIES

The annotation policies below summarize the templates used in data construction. The ALM checks the recording and transcript, retains a separate minimal grammar revision, and associates pronunciation diagnoses with the spoken text units. The checked transcript supplies canonical pronunciations; the grammar revision supplies changes in expression. The comparison covers the complete utterance, while the final feedback contains the selected corrections and practice advice.

English evidence and correction criteria. The English routes share an exam-informed system template and use distinct user messages. Reviewed expert IPA supplies the primary segmental evidence in the expert route, with audio resolving stress and ambiguous cases. In the model-prediction route, the teacher listens first and uses candidate probabilities as supporting evidence. Thus, the two routes preserve the provenance and reliability of the supplied pronunciation evidence.

A reported error must be clearly supported by the recording, fall outside accepted contextual or accent variants, and have a material effect or recur as a meaningful pattern. Relevant effects include word identity, intelligibility, listener effort, and grammatical control or speech rhythm. Uncertain evidence alone does not qualify an issue for correction. Feedback normally contains at most three errors in total across pronunciation and grammar. It may contain up to five when more than three independent issues clearly affect meaning or intelligibility. The clearest occurrence represents a repeated pattern, and consequential issues take priority over minor differences.

Natural reduction, assimilation, linking, flapping, and equivalent notation are not errors by themselves. A reduced, masked, or devoiced function word or its absence from the supplied candidates does not alone establish an omission or substitution. The criteria draw on IELTS and TOEFL pronunciation expectations but do not generate examination scores. Learner-facing comments explain the retained diagnoses without exposing internal candidate weights; practice normally consists of one to three short items. When no important issue remains, the error lists may be empty with a maintenance suggestion.

Mandarin disagreement and contextual variants. Mandarin annotation retains a language-specific disagreement rule, expressed on a 100-point candidate-weight scale: the leading weight is below 70 and the second is at least 20. Expert-route inputs preserve the three-expert counts, whereas model-route inputs carry weights quantized upstream. The prompt normally includes high-disagreement positions among the corrections, with a local exemption when the audio clearly supports the full canonical pronunciation. Separate rules account for licensed tone sandhi, neutral tones, erhua, connected-speech weakening, and acceptable accent variants; an accent variant may still warrant correction when it impairs intelligibility.

Post-processing checks high-disagreement coverage by normalizing each stored component's counts to 100, restoring missing records required by the rule, and excluding citation-tone versus expected-sandhi conflicts. This deterministic check uses stored evidence and linguistic rules rather than making another acoustic judgment. Normalization changes the weight scale, not the number or source of expert judgments.

These annotation policies construct supervision; the benchmark's fixed scoring filters are specified separately in Appendix C.1. Learner requests in Appendix B.4 contain audio and a task instruction, without the annotation transcript or pronunciation evidence supplied to the teacher.

A.3 MANDARIN: FROM EXPERT DISAGREEMENT TO A TONE CORRECTION

The following Mandarin example is a 9.90-second recording. The transcript, expert counts, ALM annotation, and coaching target refer to the same audio interval. We retain the dataset's numbered pinyin, where `v` represents *ü* in the pronunciation strings.

Recording and pronunciation evidence
Transcript: 组织专家开展健康防护咨询
Meaning: Organize experts to provide health-protection advice.
Expert judgments for "询": onset `x`: 3/3; rime `un`: 3/3; tone 4: 2/3; tone 2: 1/3.
Canonical pronunciation: `xvn2`.

ALM annotation
Recovered content: 组织专家开展健康防护咨询
Realized pronunciation at "询": `xvn4`.
Diagnosis: "该字本应读第二声（阳平），实际发音听起来像第四声（去声），调值偏高且有下降趋势。"
Grammar: no errors.

Coaching target, excerpt
"询": 标准[xvn2]，你读成[xvn4]。
"注意'询'字是第二声，发音时声音要由中音向高音扬起，像提问时的语调。"
Practice: "朗读词组：咨询[zi1 xvn2]，重点感受'询'字第二声的上扬音调。"

The tone frequencies preserve the experts' disagreement. After checking the audio, the ALM selects a falling realization for correction to the second tone. The comparison trace may retain `xvn4 (possibly xvn2)`, while the feedback explains the selected correction. Onset, rime, and tone remain separate distributions; no whole-syllable probability is obtained by multiplying them. The change from annotation `un` to rendered `vn` is a pinyin notation convention.

A.4 English: From IPA Evidence to Word-Level Feedback

The English example is a 3.62-second recording from the expert-annotation route. The ALM recovers the text from audio, aligns the pronunciation evidence with its words, and identifies the consonant substitution in *she*. The original IPA and the correction-oriented targets remain distinguishable.

Recording and pronunciation evidence
Recovered transcript: Each day she became a more vital part of him.
Expert IPA, opening words: [ˈi tʃ ˈd eɪ ˈs i ...].
Meaning of the excerpt: *Each day she became a more vital part of him.*

ALM annotation
Corrected text: Each day she became a more vital part of him.
Word: she; **canonical:** [ˈʃ iː]; **realized:** [ˈs i].
Diagnosis: "The word 'she' was pronounced with an /s/ sound instead of the /ʃ/ sound."
Grammar: no errors.

Coaching target, excerpt
"she": target [ʃˈiː] vs. you said [sˈi].
"Use the /ʃ/ sound, like in 'shoe' or 'ship'."
Practice: she [ʃˈiː], shoe [ʃˈuː], ship [ʃˈɪp].

The annotation record retains differences that do not become corrections. For example, the initial vowel in *Each* has no length mark in the original IPA. When rendering the training target, uncorrected words use canonical forms and *she* retains the realization attached to its correction. This target represents the selected diagnosis; the original pronunciation evidence remains unchanged. Appendix B.2 shows how the same diagnosis supports Thinking and Instruct responses without repeating the annotation step.

B Training Tasks and Response Formats

B.1 Task Coverage and Mixture

The training collection contains approximately 20k instances: 13k single-turn and 7k multi-turn examples, around 15k unique recordings with equal numbers of Mandarin and English correction tasks. These are instance counts; several targets may share a recording. Table 5 lists the tasks supported by the annotations. Each takes audio and a task request, without a learner-supplied reference transcript. Transcription and pronunciation prediction supervise intermediate analysis, while diagnosis and coaching combine these steps. Accent description is restricted to Mandarin, yielding eight Mandarin and seven English categories. Sampling assigns comparable quotas across the available tasks within each language and approximately equal mass to the two audio languages.

Requests may be written in Mandarin or English; the phonetic representation follows the recording language. The coaching mixture combines single-turn and multi-turn examples in Thinking inline and Instruct formats at a 1:1 ratio, using the same pronunciation annotations and correction targets.

Task	Mandarin	English	Target
Transcription	Yes	Yes	Transcript
Accent description	Yes	No	Accent category
Grammar checking	Yes	Yes	Corrections & explanations
Canonical pronunciation	Yes	Yes	Pinyin / IPA
Realized pronunciation	Yes	Yes	Realized sequence
Error detection	Yes	Yes	Contrasts & corrections
Structured diagnosis	Yes	Yes	JSON diagnosis & practice
Coaching feedback	Yes	Yes	Feedback & practice

Table 5: Training-task coverage. Pronunciation-error detection, structured diagnosis, and complete coaching feedback support explicit comparison traces and direct answers; the other tasks supervise their respective outputs.

B.2 Response Modes and Serialization

Thinking responses expose pronunciation comparisons and correction reasoning before feedback; Instruct responses directly provide diagnoses, comments, and practice plans. Thinking inline places the comparison in the visible answer without a separate reasoning block. The comparison can also be serialized inside `<think>` and `</think>` (Table 6). This is a serialization choice, not a third response mode: a block alone does not determine visibility to the learner, and an empty block contains no reasoning. These forms reuse the same diagnosis rather than requesting another ALM judgment.

Serialization	Assistant training target
Reasoning block	`<think>` Comparison trace `</think>` + feedback
Thinking inline	Comparison trace + feedback in the visible answer
Instruct	Feedback only

Table 6: Serializations of a shared diagnosis. The first two retain the comparison trace in a separate block or inline; Instruct retains final feedback. Empty format-control blocks are omitted.

The Mandarin and English examples in Appendix A supply concrete diagnoses and feedback for these formats. Thinking exposes the associated comparison trace; Instruct retains the final correction and practice guidance. The examples illustrate training targets, not benchmark predictions.

B.3 Training Comparison Conditions

The *Main* condition uses pronunciation probabilities for ALM annotation, mixes Thinking and Instruct supervision, and is evaluated with Instruct responses. *Discrete targets* replaces those pronunciation probabilities in the annotation input with one realized pronunciation, so the supervision may differ for the same recording; this is not an output-format-only comparison. *w/o Thinking data* removes Thinking supervision while retaining Instruct final feedback; initialization, underlying data, sample count, training steps, and decoding settings otherwise match Main. *Scale* expands balanced training data while preserving equal Mandarin/English sampling and task/response-form proportions. All conditions share evaluation references, filters, localization, and metrics. Candidate agreement uses shared normalization and does not assess calibration of pronunciation probabilities.

Pronunciation Probabilities in the Output. Pronunciation probabilities may inform annotation without appearing numerically in feedback, as the Mandarin example in Appendix A.3 illustrates. Discrete targets changes this annotation evidence, rather than merely serializing a fixed diagnosis differently. Explicitly requesting candidate values and weights is a separate output requirement, even for the same recording. The numerical-output format concerns Mandarin realized-pronunciation prediction, structured diagnosis, and complete coaching feedback, leaving the other five task formats unchanged; it is not a separate reported comparison. Such probability strings are learned through autoregressive token likelihood, rather than the distributional cross-entropy used to train HuPER.

B.4 Representative Learner Requests

The following requests are taken from the training pools. Each accompanies audio without a reference transcript or supplied pronunciation sequence. Thinking requests explicit intermediate comparisons; Instruct requests the final diagnosis and guidance. The paired languages express the same task, and practice remains part of complete coaching rather than a separate supervised task. All eight requests retain their original wording. This is only an example, we have more paraphrased prompt for training and additional unseen prompt for test set.

Instruct: coaching with practice. *Chinese.* 请全面点评我的发音和口语：先写出我说的内容；检查语法有没有错误；逐个指出发音问题并给出改法；最后给我一份发音练习计划。

English. Give me full feedback on my speaking, covering: the transcript of what I said, any grammar errors, each pronunciation issue with how to correct it, and a closing practice plan.

Thinking: staged coaching. *Chinese.* 请分步骤全面评估我的发音和口语：第一步判断语种并逐字转写；第二步给出整句标准读音和实际读音；第三步逐字对比找出差异、确定发音错误；第四步检查语法错误；第五步逐个分析发音问题并给出改法；最后制定发音练习计划。

English. Walk through my speech in stages: determine the language and give a verbatim transcript; lay out the full target pronunciation and my actual pronunciation; compare them word by word and sound by sound to confirm the errors; review my grammar; analyze every pronunciation issue with corrections; and finish with a standard-pronunciation practice plan.

Instruct: structured diagnosis. *Chinese.* 请直接给出我的发音诊断结果，以结构化JSON 输出，字段为transcription（字符串）、corrected_text（字符串）、grammar_errors（数组，每项含heard, corrected, type, how_to_fix，均为字符串）、pronunciation_errors（数组，每项含unit, target, heard, type, how_to_fix，均为字符串）、practice_plan（字符串数组）、feedback（字符串）。

English. Please give your pronunciation diagnosis directly as structured JSON with these keys: transcription (string), corrected_text (string), grammar_errors (array of objects with heard, corrected, type, how_to_fix — all strings), pronunciation_errors (array of objects with unit, target, heard, type, how_to_fix — all strings), practice_plan (array of strings), feedback (string).

Instruct: pronunciation diagnosis. *Chinese.* 请分析我整句话中的每一个发音问题，逐个分析和对比，并给出改法。

English. Analyze every pronunciation problem in my sentence one by one, compare each with the correct pronunciation, and give me the fixes.

B.5 Diagnosis Reward

The GRPO reward is computed from the final structured response to constrain its format and reduce missed corrections in Mandarin and English. Each reported correction is compared with the reference position, canonical pronunciation, and realized pronunciation. Matching position and pronunciation earns full credit; a position-only match earns half credit.

Mandarin human-reference positions have weight 1 when experts agree on all pronunciation components and 0.5 otherwise. Matching uses a deterministic reference for each component; the weight reflects expert agreement. HuPER-only deviations can earn supplementary credit with weight 0.5, but missing them does not lower human-reference recall.

Let C_{pred} denote credited predictions, including permitted HuPER credit; C_{human}, credit at human-reference positions; W_{human}, their total reference weight; and L_{report}, all reports, including duplicates and unlocalized corrections. Training precision and recall are

$$P_{\text{tr}} = \frac{C_{\text{pred}}}{L_{\text{report}}}, \qquad R_{\text{tr}} = \frac{C_{\text{human}}}{W_{\text{human}}}. \tag{2}$$

They are combined with the existing recall emphasis:

$$F_\beta = \frac{(1+\beta^2)P_{\text{tr}}R_{\text{tr}}}{\beta^2 P_{\text{tr}} + R_{\text{tr}}}, \qquad \beta = 1.2. \tag{3}$$

If the F_β denominator is zero, $F_\beta = 0$. Recall is 1 when the human reference contains no deviations. With no reported corrections, precision is 1 if the human-reference set is also empty, and 0 otherwise.

The indicator I_{parse} is 1 when the final answer parses as a JSON object and 0 otherwise. The indicator $I_{\text{canonical}}$ is 1 only when the complete canonical pronunciation sequence matches the reference after normalization, and 0 otherwise. Equation 1, including its parsing prerequisite, can therefore be written as

$$\mathcal{R} = \begin{cases} 0, & I_{\text{parse}} = 0, \\ 0.85F_{1.2} + 0.10 + 0.05I_{\text{canonical}}, & I_{\text{parse}} = 1. \end{cases} \tag{4}$$

GRPO uses multiple sampled responses per input, group-normalized reward advantages, and KL regularization against a reference policy. It optimizes structured diagnosis; SFT supplies supervision for explanations and practice guidance.

C Evaluation Protocol and Metrics

C.1 Reference Coverage and Correction Filters

The Diagnostic Track of PronCoach-Bench contains only recordings with complete, localizable coaching references: 296 Mandarin and 214 English utterances. All 510 recordings enter scoring. Human evidence is stored separately from diagnostic references: Mandarin retains independent hearing trajectories and candidate frequencies from three experts, while English retains quality-controlled final IPA transcriptions. An ALM conditions on the recording, transcript, and human realized pronunciations to generate position-linked text units, target pronunciations, realized pronunciations, and error types. These diagnoses are not expert-adjudicated corrective advice. Eligibility depends on reference availability and localizability; missing or unparseable model responses do not change the reference set.

The Mandarin filter lists modal particles, structural particles, and prepositions, and excludes a reference correction only when its character unit matches the list. English filtering excludes listed function words. The other two English rules normalize target and heard forms, then check for exactly one specified voicing substitution, or exactly one specified short-vowel substitution in a word of at most four letters. These rules do not exclude words with additional differences or remove the whole recording.

Mandarin has 456 original reference correction entries; the word-list filter removes 92, leaving 364 positions. English has 245 original entries, from which 38 function-word, 14 voicing, and four short-vowel entries are removed. Consolidating 13 repeated entries at retained positions leaves 176 positions. For recording u, let S_u denote excluded positions and G_u the retained reference correction set. Predictions on S_u are also removed from scoring; other positions without a reference exclusion remain eligible for false corrections.

Table 7: Counts for PronCoach-Bench. Diagnostic correction positions are counted after filtering and consolidation. Each Clean-Speech recording is evaluated once with a request in the corresponding language. The supplementary PER corpora are not included in these counts.

Track	Audio language	Recordings	Runs	Targets
Diagnostic	Mandarin	296	296	364
Diagnostic	English	214	214	176
Clean-Speech	Mandarin	81	81	0
Clean-Speech	English	97	97	0
Total	–	688	688	540

C.2 System Inputs and Response Handling

At evaluation, Main and all three training variants are reference-free coaches that produce Instruct responses from a recording and language-appropriate correction request, without a reference transcript, diagnostic reference, or scoring filter. The *w/o Thinking data* condition names a training ablation, not a different evaluation response format. The public audio-language baselines including

Doubao Seed 2.0 Lite, Gemini 3.8 Flash, Gemini 3.1 Pro Preview, Muse Spark 1.2, and Qwen3.5-Omni-plus receive the same reference-free input type but answer in free-form coaching text. For reference-assisted phonetic systems, a deterministic adapter aligns recognized phones with supplied canonical pronunciations to produce word-level corrections; this describes the comparison pipeline and does not imply that every recognizer itself requires canonical pronunciations. Recordings, retained references, filters, normalization, and scoring remain shared within each reported scope.

Native structured diagnostic responses are parsed directly. Free-form coaching responses are converted to the same correction-entry interface by the fixed, temperature-zero GPT-4o-mini extractor described in Appendix C.4; its input is the evaluated model's answer alone. It receives neither audio nor any reference. Missing responses remain in the scheduled evaluation set and follow the failure rules above.

C.3 Reference-Assisted Phonetic Baselines

The reference-assisted comparison uses CROTTC-IF (Geng et al., 2026), MDD-LLM (Llama-3.2)[6], and MPL-MDD (Yang et al., 2022) on 214 English Diagnostic and 97 English Clean-Speech recordings. Their adapter aligns recognized phones with supplied canonical pronunciations; fixed PronCoach transcriptions provide Clean-Speech reference text, without generated correction or realized-pronunciation fields. The resulting entries use the same English references, normalization, filters, localization, and scorer as the reference-free systems, while the English aggregate remains distinct from the bilingual score. W2V2-eSpeak (Xu et al., 2021), ZIPA-CR-Large, and ZIPA-CR-NS-Large (Zhu et al., 2025) are additionally compared through PER. Missing coaching cells denote unreported results, not zeros or evidence that a system cannot perform correction.

C.4 Answer Interpretation and Localization

The language-model extractor maps a coaching answer to its stated text units, target pronunciations, realized pronunciations, and error types without reading references or audio. It preserves candidate alternatives in Mandarin descriptions such as being between two sounds. For English answers describing only a changed phone, the extraction prompt reconstructs a whole-word realized pronunciation by applying the stated change to the target pronunciation in the answer; if reconstruction is impossible, it retains the stated sound. This step interprets the answer's pronunciation claim rather than recognizing the recording again.

Grammar advice, practice plans, encouragement, and general feedback do not create pronunciation-correction entries. An explicit statement that no correction is needed produces an empty list. Native structured output is parsed directly; other answers use the same fixed GPT-4o-mini extraction prompt at temperature zero. The answer-only input prevents the extractor from completing a diagnosis using diagnostic references, although English word-level reconstruction can still introduce extraction error.

Localization first matches normalized text units and target pronunciations, with a surface-only fallback when target matching fails. Repeated characters or words are assigned monotonically in response order. Multiple solutions use a deterministic earliest assignment; when no monotone solution exists, a deterministic greedy assignment is used. Complete correction entries that cannot be placed in the reference text are counted separately. Multiple entries mapped to one position contribute one position record rather than repeated hits.

Realized-pronunciation matching uses a common notation normalization that removes formatting delimiters, stress and length marks, and folds specified rhotic and syllabic variants. Mandarin compares candidate-set overlap; English compares one whole-word realized pronunciation. This matching does not separately assess stress or duration. Correct localization likewise does not establish the correctness of a proposed target pronunciation or articulatory explanation.

C.5 Diagnostic, Clean-Speech, and Aggregate Metrics

The metrics below correspond to the result-table columns. Higher values are better except for false-correction density (FC) and the supplementary phoneme error rate (PER; Appendix C.6), for which lower values are better. We use $\mathbf{1}[\cdot]$ for an indicator.

[6] https://huggingface.co/Haopeng/MDD-LLM-Llama3.2-1B-L2-ARCTIC

C.5.1 SCORING UNITS AND NOTATION

Diagnostic scoring uses character positions for Mandarin and word positions for English. For recording u, let U_u be the reference positions, S_u the filtered positions, $G_u \subseteq U_u \setminus S_u$ the reference correction targets, and $D_u \subseteq U_u \setminus S_u$ the localized predicted corrections. Repeated predictions at one position count once, with the last retained entry used for realized-pronunciation matching. Let E_u count complete correction entries that cannot be localized in the reference text.

Counts are pooled across recordings within each language before computing ratios. Here G counts reference targets, T matched positions, P localized predictions, F predictions at non-target positions, E out-of-text entries, and N unfiltered non-target positions:

$$G = \sum_u |G_u|, \qquad T = \sum_u |D_u \cap G_u|, \qquad P = \sum_u |D_u|,$$
$$F = \sum_u |D_u \setminus G_u|, \qquad E = \sum_u E_u, \qquad N = \sum_u |U_u \setminus (G_u \cup S_u)|. \tag{5}$$

Let $M_{u,i}$ indicate realized-pronunciation agreement at a matched position. Mandarin requires overlap between normalized predicted and reference candidate sets; English requires equality of normalized whole-word realized pronunciations. The notation normalization above removes stress and length marks. The count

$$H = \sum_u \sum_{i \in D_u \cap G_u} M_{u,i} \tag{6}$$

measures realized-pronunciation agreement among hits; it does not independently assess the proposed target pronunciation, articulatory guidance, or calibration of pronunciation probabilities.

C.5.2 DIAGNOSTIC METRICS

Recall R measures coverage of reference correction targets. *Localization precision* P_{loc} measures the fraction of localized predictions that identify a reference target. *Realized-pronunciation agreement* $\mathrm{ACC}_{\mathrm{hear}}$ measures how often the stated realized pronunciation matches the reference, conditional on correct localization:

$$R = \frac{T}{G}, \qquad P_{\mathrm{loc}} = \frac{T}{P}, \qquad \mathrm{ACC}_{\mathrm{hear}} = \frac{H}{T}. \tag{7}$$

For $\ell \in \{\mathrm{zh}, \mathrm{en}\}$, *localization F1* balances target coverage with localization precision, while *realized-pronunciation F1* balances coverage with realized-pronunciation agreement among hits:

$$F1_\ell^{\mathrm{loc}} = \frac{2R_\ell P_{\mathrm{loc},\ell}}{R_\ell + P_{\mathrm{loc},\ell}}, \qquad F1_\ell^{\mathrm{hear}} = \frac{2R_\ell \mathrm{ACC}_{\mathrm{hear},\ell}}{R_\ell + \mathrm{ACC}_{\mathrm{hear},\ell}}. \tag{8}$$

Because $\mathrm{ACC}_{\mathrm{hear}}$ is conditional agreement rather than precision, $F1_\ell^{\mathrm{hear}}$ is a benchmark-specific harmonic summary, not conventional precision–recall F1.

False-correction density measures erroneous interventions per unfiltered non-target position:

$$\mathrm{FC} = \frac{F + E}{N}. \tag{9}$$

Lower is better. Out-of-text entries contribute to FC but not to P or P_{loc}. Consequently, FC can exceed one and is not a bounded error probability.

C.5.3 CLEAN-SPEECH METRICS

Clean-Speech accuracy is computed separately for each language. The completed evaluation contains $n_{C,\mathrm{zh}} = 81$ Mandarin and $n_{C,\mathrm{en}} = 97$ English recordings (37 IELTS and 60 TOEFL), each evaluated once with a correction request in the corresponding language. Let v_u indicate a valid response, p_u count its pronunciation corrections, and g_u indicate whether it reports a grammar issue. For each language ℓ,

$$K_\ell = \sum_{u=1}^{n_{C,\ell}} \mathbf{1}[v_u = 1,\ p_u = 0,\ g_u = 0], \qquad \mathrm{ACC}_{c,\ell} = \frac{K_\ell}{n_{C,\ell}}. \tag{10}$$

A structured answer is clean only when both correction lists are valid and empty; an extracted answer must contain no pronunciation correction and no grammar issue. A clean flag cannot override an explicit correction. The tables report $\mathrm{ACC}_{c,\ell}$ by language. For bilingual aggregation, we pool the clean-response counts before applying the benchmark's monotone transformation:

$$\overline{\mathrm{ACC}}_c = \frac{K_{\mathrm{zh}} + K_{\mathrm{en}}}{178} = \frac{81\,\mathrm{ACC}_{c,\mathrm{zh}} + 97\,\mathrm{ACC}_{c,\mathrm{en}}}{178}, \qquad F1_C = \frac{2\overline{\mathrm{ACC}}_c}{1 + \overline{\mathrm{ACC}}_c}. \tag{11}$$

This score preserves the ordering of pooled clean accuracy; it is not a conventional two-class F1 estimated on the all-clean set. The English-only comparison instead uses $F1_{C,\mathrm{en}} = 2\,\mathrm{ACC}_{c,\mathrm{en}}/(1+\mathrm{ACC}_{c,\mathrm{en}})$.

C.5.4 Aggregate scores

For $m \in \{\mathrm{loc}, \mathrm{hear}\}$, ***Diagnostic F1*** averages the language-specific summaries using reference-target counts, $n_{\mathrm{zh}} = 364$ and $n_{\mathrm{en}} = 176$:

$$F1^m_D = \frac{n_{\mathrm{zh}} F1^m_{\mathrm{zh}} + n_{\mathrm{en}} F1^m_{\mathrm{en}}}{n_{\mathrm{zh}} + n_{\mathrm{en}}} = \frac{364 F1^m_{\mathrm{zh}} + 176 F1^m_{\mathrm{en}}}{540}. \tag{12}$$

Overall F1 combines diagnosis and performance on clean speech using $n_D = 510$ and $n_C = 178$ recordings:

$$F1^m_{\mathrm{all}} = \frac{n_D F1^m_D + n_C F1_C}{n_D + n_C} = \frac{510 F1^m_D + 178 F1_C}{688}. \tag{13}$$

It averages the component scores rather than computing a micro-F1 from pooled cross-track counts.

For the English-only comparison, ***English aggregate F1*** uses $n_{D,\mathrm{en}} = 214$ Diagnostic recordings and $n_{C,\mathrm{en}} = 97$ Clean-Speech recordings:

$$F1^m_{\mathrm{E}} = \frac{n_{D,\mathrm{en}} F1^m_{\mathrm{en}} + n_{C,\mathrm{en}} F1_{C,\mathrm{en}}}{n_{D,\mathrm{en}} + n_{C,\mathrm{en}}} = \frac{214 F1^m_{\mathrm{en}} + 97 F1_{C,\mathrm{en}}}{311}. \tag{14}$$

Its scope differs from the bilingual overall score.

C.5.5 Boundary cases and supplementary utterance-level measure

A ratio with a zero denominator remains undefined, as do summaries requiring it. When both harmonic-mean inputs are zero, the summary is defined as zero; otherwise the stated formula applies to defined inputs. Missing or failed Diagnostic answers supply no predictions while retaining their reference targets and non-target positions. Missing or empty Clean-Speech answers and parsing or extraction failures are non-clean and remain in the denominator. Reaching the generation limit alone does not invalidate successfully parsed or extracted content.

The free-form adapter omits entries without a usable text unit or with both pronunciation fields empty. For Diagnostic recordings without a retained reference target, the supplementary *utterance-level false-correction rate* is

$$\mathrm{FCR}_{\mathrm{clean}} = \frac{\sum_u \mathbf{1}[G_u = \varnothing]\,\mathbf{1}[|D_u| + E_u > 0]}{\sum_u \mathbf{1}[G_u = \varnothing]}. \tag{15}$$

This measures whether any scored pronunciation correction occurs on such a recording. It differs from table FC and from the separate Clean-Speech rate ACC_c. Having no retained target does not imply that every acoustic variation is canonical.

C.6 Supplementary Phoneme Error Rate Evaluation

The supplementary PER evaluation uses test-set recordings from L2-ARCTIC (Zhao et al., 2018), SpeechOcean762 (Zhang et al., 2021), and EpaDB (Vidal et al., 2019), separately from the 688-recording two-track benchmark. For PronCoach, the evaluated phones are extracted from its complete coaching answer. The extractor selects the stated realized-pronunciation sequence in the final answer, excluding the reasoning block, canonical pronunciations, and isolated contrasts in explanatory text. Conflicting realized-pronunciation sequences are treated as extraction failures rather

than selecting the sequence with the best reference match. HuPER predictions are not substituted for the answer.

Reference and predicted sequences are normalized to a common representation built from 39 ARPABET phones and mapped IPA symbols. ARPABET stress digits and silence or control tokens are removed, and `AX` maps to `AH`. IPA is Unicode-decomposed, with combining diacritics, stress, length, and aspiration marks removed under the broad-transcription convention. Equivalent symbol spellings are normalized, syllabic consonants expand to a schwa followed by the consonant, and documented broad variants such as rhotic vowels map to the corresponding shared tokens. Longest-match tokenization is confined to each explicitly delimited chunk, so phones are not joined across word or notation boundaries.

For each corpus, reference and predicted sequences are aligned by minimum edit distance. If S_u, D_u, and I_u count substitutions, deletions, and insertions, and N_u is the number of reference phones, then

$$\mathrm{PER}\,(\%) = 100\,\frac{\sum_u (S_u + D_u + I_u)}{\sum_u N_u}. \tag{16}$$

Lower is better, denoted by $\mathrm{PER}\downarrow$; this is corpus-level micro-PER, not the mean of utterance-level percentages. Eligibility is determined from the references alone: recordings with indeterminate `ERR` labels, unmappable reference symbols, or empty references are excluded consistently for all systems. Unknown predicted phones retain an `<unk>` token. Missing, unparseable, or truncated outputs are scored as empty predictions on the eligible set, contributing N_u deletions without reducing the denominator; parsing coverage is recorded separately. This truncation rule is specific to PER and differs from the content-based Diagnostic and Clean-Speech rules above.

C.7 Auxiliary HuPER Perception Protocol

The auxiliary HuPER analysis uses 300 Mandarin recordings: 200 contain at least one difference between the expert-plurality pronunciation and the canonical pronunciation, and 100 contain none. The subset comprises 4,308 characters and 12,924 onset, rime, and tone positions. Canonical pronunciations are generated with pypinyin; the realized-pronunciation reference is the plurality of three independent expert judgments. Ties favor the canonical candidate when present and otherwise follow lexical order.

The positive class denotes a canonical-pronunciation mismatch before a contextual decision about correction. Selection deliberately enriches mismatch-containing recordings. This protocol therefore concerns component-level phonetic perception, separately from both PronCoach-Bench tracks and the supplementary English PER evaluation.

D SUPPLEMENTARY RESULTS

Table 8 complements the main results in Table 1 with recall, localization precision, the pooled Clean-Speech transformation, and localization-based overall F1. The predictions, recordings, references, filters, and scorer are unchanged. The first two metrics are language-specific; the aggregate columns span both languages and follow Equation 13.

Table 8: Complementary metrics for the reference-free systems. $F1_C$ pools both Clean-Speech languages, and $F1^{\text{loc}}_{\text{all}}$ combines localization performance with clean-speech behavior. Realized-pronunciation agreement, localization F1, FC, and realization-based overall F1 are reported in the main table.

Model	Language	Diagnostic		Clean-Speech	Overall
		$R \uparrow$	$P_{\text{loc}} \uparrow$	$F1_C \uparrow$	$F1^{\text{loc}}_{\text{all}} \uparrow$
Doubao Seed 2.0 Lite	Mandarin	0.4011	0.1281	0.0546	0.1648
	English	0.4318	0.1496		
Gemini 3.8 Flash	Mandarin	0.2967	0.0942	0.5042	0.2528
	English	0.5341	0.1313		
Gemini 3.1 Pro Preview	Mandarin	0.3654	0.0805	0.6716	0.2919
	English	0.2443	0.1937		
Muse Spark 1.2	Mandarin	0.2857	0.0778	0.4178	0.2112
	English	0.5852	0.1020		
Qwen3.5-Omni-plus	Mandarin	0.4698	0.1377	0.4721	0.2724
	English	0.3977	0.1176		
PRONCOACH	Mandarin	0.8187	0.1930	0.5528	0.4383
	English	0.5909	0.5622		