FuzhouBench: A Multimodal Benchmark for Eastern Min Chinese with Tone Sandhi Annotations

Anonymous ACL submission

Abstract

Eastern Min Chinese (Fuzhou dialect, ISO 639-3 cdo) has roughly ten million speakers yet no dedicated paper at any major NLP venue. Four properties make Fuzhou structurally distinct from Mandarin and challenging for current NLP: (1) a tone-sandhi system conditioned on the following syllable's citation tone, (2) initial-consonant assimilation in non-final positions, (3) systematic literary–colloquial diglossia (45% of frequent characters carry $\geq$2 distinct segmental readings), and (4) substantial lexical divergence (everyday concepts use entirely different morphemes from Mandarin). We release **FuzhouBench**, the first public NLP benchmark for Eastern Min, built on a community-maintained open dictionary: 22,890 word entries with 51,505 standardized Yngping romanizations, 11,276 explicit tone-sandhi annotations, 5,160 explanations, and 3.06 hours of paired dictionary-style audio from two native speakers. We define three tasks (grapheme-to-phoneme, sandhi prediction, definition retrieval) with full baselines. Frontier LLMs score under 3% word-exact on Yngping G2P; a 62-entry symbolic sandhi rule beats every LLM by 7–13 points; and BM25 collapses from MRR 0.799 to 0.372 when queries shift from dictionary-extracted to colloquial Mandarin, partially recoverable via a two-stage hybrid + cross-encoder pipeline (MRR 0.716). **Supplementary materials include the FuzhouBench dataset with 10 audio samples, and baseline code** (CC-BY-NC-SA 4.0 for data, MIT for code); an anonymized snapshot accompanies this submission.

1 Introduction

Chinese is a big umbrella term in linguistic reality, as it is a family of varieties: Mandarin, Cantonese, Eastern Min, Hakka, Hokkien, etc. are not mutually intelligible with one another. Section 2 gives a brief linguistic profile of Fuzhou Eastern Min and positions it within Sinitic.

NLP coverage of these varieties is highly uneven, mirroring the broader linguistic-diversity gap documented by Joshi et al. (2020) and the survey of low-resource methods by Hedderich et al. (2021); Bird (2022) reframes the resource scale to argue that local-language scenarios deserve their own treatment rather than being subsumed under generic "low-resource" work. Cantonese has received the bulk of recent attention through resources such as Yue-Benchmark (Jiang et al., 2025b) and Yue-Data/YueTung (Jiang et al., 2025a). Taiwanese Hokkien has at least one EMNLP paper on singing voice synthesis (Bai et al., 2023) and a code-mixing corpus (Lu et al., 2022). Other major Sinitic varieties remain almost invisible: to the best of our knowledge, no NLP paper at a major venue has taken Eastern Min as its primary subject.

This paper releases the first public NLP benchmark for Eastern Min, focusing on the Fuzhou variety. The data builds on an open, community-maintained Fuzhou dialect dictionary that has been curated by volunteer native speakers for several years.[1] The source dictionary already structures three traditional references: contemporary lexical entries, the *Fuzhou Dialect Dictionary* (Feng and Li, 1998) (a modern Chinese-language consolidation of earlier references including Maclay (1870); the maintainers label these tables "Feng" after the editor's surname, and we keep that label), and the *Qi-Lin Eight-Tone* rhyme book. The most striking baseline finding is on definition retrieval: BM25 looks strong on dictionary-original queries (MRR 0.799) but collapses to MRR 0.372 when queries are replaced with LLM-generated colloquial Mandarin, and a two-stage hybrid + cross-encoder reranker is needed to recover (MRR 0.716). This echoes

[1] Some of the authors are involved with the source community. We follow ACL anonymity practice by not naming individual contributors here; the source dictionary's name, contributor identities, and affiliations are disclosed in the camera-ready version.

the dialect gap that Kantharuban et al. (2023) report at the language level: query distribution dominates retrieval performance for low-resource varieties. The cross-encoder scores a (query, document) pair jointly, producing a finer signal than the separate query and document vectors used in first-stage retrieval; we apply it to the top-50 hybrid candidates and report top-10 metrics.

Our contributions:

1. A community-maintained v1.0 release of the first public Eastern Min multimodal NLP benchmark, comprising 22,890 word entries, 51,505 Yngping readings (standardized Fuzhounese romanizations), 11,276 entries with explicit tone-sandhi annotations, 5,160 explanations, and 3.06 hours of dictionary-style audio from two native speakers.

2. Definitions of three benchmark tasks suited to the data: grapheme-to-phoneme (G2P), tone-sandhi prediction, and definition retrieval. Each with complete baselines spanning open Qwen3 models (1.7B, 8B, 30B-A3B), the closed `gpt-5.4-mini`, symbolic rule-based predictors for sandhi, and a two-stage hybrid + cross-encoder pipeline for retrieval.

3. An empirical analysis of where current LLMs fail on Fuzhou: Mandarin-Pinyin-style output instead of Yngping on G2P, inability to apply symbolic sandhi rules even with pattern-matched 5-shot demonstrations, and BM25 collapse on colloquial Mandarin queries that only a two-stage retrieval pipeline can recover from.

2 Background: Fuzhou Eastern Min

Family and speakers. Fuzhou is the representative variety of Eastern Min, a branch of the Min subfamily of Sinitic (Figure 1). Min sits alongside Mandarin, Cantonese (Yue), Wu, Hakka, Xiang, and Gan as the major Sinitic branches; Min itself further subdivides into Southern Min (Hokkien, Teochew), Eastern Min, Min Bei, and smaller groups, none of which are mutually intelligible with each other or with the other Sinitic branches. Eastern Min has roughly ten million speakers, concentrated in the city of Fuzhou in eastern Fujian province and in diaspora communities across Southeast Asia and North America, well below Hokkien (around 50 million), Mandarin (around 1.4 billion) or Cantonese (around 85 million) (Eberhard et al., 2024).

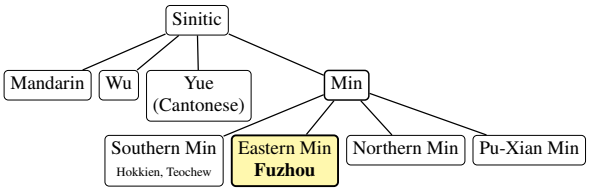

Figure 1: Position of Fuzhou Eastern Min within Sinitic. This figure is not a comprehensive Sinitic classification (Hakka, Gan, Xiang, and several smaller branches are omitted, as are finer subdivisions inside Mandarin, Wu, and Yue); its only purpose is to show readers where Fuzhou sits relative to Mandarin, Cantonese, and the other Min varieties. The Min sub-branches are not mutually intelligible with one another or with the rest of Sinitic. FuzhouBench targets the highlighted node.

Tones and sandhi. Citation phonology distinguishes nine surface contours (55, 53, 33, 21, 24, 242, 213, 5, 0; the digits are Chao (1930)'s five-level pitch notation, where 1 is lowest and 5 is highest, and the full key is in Appendix B), against Mandarin's four and Cantonese's six. Non-final syllables systematically change tone based on the following syllable's citation tone (Section 6.2); the full mapping spans roughly sixty (curr, next) contexts, against Mandarin's single dominant rule (third-tone sandhi).

Initial-consonant assimilation. A feature uncommon in Sinitic: citation s, d, g surface as l, l, $\varnothing$ in non-final positions depending on the preceding tone and segment. The phenomenon is reported for Eastern Min and a few Min Bei varieties; it is absent from Mandarin, Cantonese, and most of Hokkien.

Literary–colloquial diglossia and orthography. Characters carry distinct literary (*wen*) and colloquial (*bai*) readings, with broader coverage than in Mandarin: many morphemes take one reading in formal compounds and another in casual single-character use. Empirically, of the 2,893 unique characters observed in the dictionary's multi-character compound citations, 18% (519) show $\geq$2 distinct segmental readings (tone digits collapsed, so sandhi is excluded); the rate rises to 45% (232) among the 515 unique characters that appear in 10+ different compounds.[2] Many native morphemes also lack a standard Mandarin character and require dialect-specific characters or romanizations; conventions in current use include Yngping romanization (which

[2]Reproduced by `diglossia_stats.py` in the released code. The 10+ frequency control prevents rarely-attested characters from depressing the rate.

Phenomenon	Mandarin	Fuzhou Eastern Min
Tones and sandhi	4 contour tones; one dominant sandhi rule (3rd-tone).	9 surface contours (7 citation + 5 sandhi); sandhi rule conditioned on the next syllable's citation tone.
Ex. "Fu-zhou" = the city Fuzhou	Fú zhōu (no sandhi)	**houk24** ziu55 → **huk21** ziu55
Initial-consonant assimilation	Absent.	*s, d, g* → *l, l,* ∅ in non-final positions.
Ex. "two-ten" = twenty	èr shí (no assimilation)	**nei242** **s**eik5 → **ni55** **l**eik5
Literary–colloquial diglossia	Sporadic; mostly one reading per character.	Systematic; many characters carry two readings (formal compounds take the literary reading, casual the colloquial).
Ex. character "big" in two compounds	dà gài "big-summary" = approximately dà tóu "big-head" = blockhead	**dai**53 kai213 (literary "big") **duai**55 lau53 (colloquial "big")
Lexical divergence from Mandarin	Standard Mandarin morphemes.	Min-specific morphemes for many everyday concepts (Appendix C catalogs 14 pairs).
Ex. "go home"	回家 huí jiā (lit. "return-home")	轉厝 duong55 cuo213 (lit. "return-house")

Table 1: Mandarin vs. Fuzhou Eastern Min on four phenomena that shape the benchmark tasks (Section 2). Fuzhou forms are in Yngping; → marks the citation-to-sandhi mapping (where applicable). **Bold blue** marks parts that change under sandhi (tone and/or vowel) or that differ between literary and colloquial readings; **bold red, underlined** marks initial-consonant assimilation. "Two-ten" shows both mechanisms acting on the same word.

we adopt). The absence of a single canonical orthography itself complicates data collection and tokenization.

Lexical divergence from Mandarin. Beyond different pronunciations of shared characters, many basic concepts use entirely different morphemes in Fuzhou: "pot", "chopsticks", "house", "stand", and "egg" share no character with their Mandarin counterparts, and verb–object collocations diverge similarly (Appendix C lists 14 pairs). This lexical gap is independent of the phonological alternations above and is the structural reason BM25 collapses on colloquial-Mandarin queries (Section 6.3).

3 Related Resources

Cantonese resources. Yue-Benchmark (Jiang et al., 2025b) translates several MMLU-style benchmarks into Cantonese and reports that mainstream LLMs lag their Mandarin performance by a substantial margin. YueData and YueTung (Jiang et al., 2025a) push this further by collecting two billion Cantonese tokens and producing a domain-adapted 7B model. CompLex-ZH (Qiu et al., 2024) provides Mandarin and Cantonese lexical-complexity annotations. Cantonese NLP surveys (Xiang et al., 2024) catalog the state of the field.

Other Sinitic varieties. Taiwanese Hokkien has a code-mixing corpus (Lu et al., 2022) and a singing voice synthesis dataset around Gezi opera (Bai et al., 2023). Tone2Vec (Yang et al., 2024) introduces pitch-based tone representations and clusters dialect transcriptions across Sino-Tibetan varieties. TWBias (Hsieh et al., 2024) benchmarks bias in Traditional Chinese LLMs through a Taiwanese cultural lens. Wu, Hakka, Xiang, and the various Min varieties beyond Hokkien (including Eastern Min) lack any comparable public NLP resource; Liu and Best (2025) survey the Sino-Tibetan NLP literature and reach the same conclusion at family scale.

4 Data Collection

4.1 Source

The source is an open online dictionary maintained by volunteer native speakers of Fuzhou Eastern Min. The project inherits content from three references. First, contemporary lexical entries collected from native speakers and printed reference works. Second, the *Fuzhou Dialect Dictionary* (Feng and Li, 1998), a modern Chinese-language consolidation that organized and translated earlier references, including the canonical nineteenth-century *Alphabetic Dictionary of the Chinese Language in the Foochow Dialect* (Maclay, 1870), into a single annotated lexicon; the maintainers label the tables derived from Feng and Li (1998) as "Feng" (after the editor's surname), and we keep that label throughout. Third, the *Qi-Lin Eight-Tone* rhyme book (Qi-Lin Ba-Yin), an eighteenth-century rhyme dictionary whose categorical labels we keep as his-

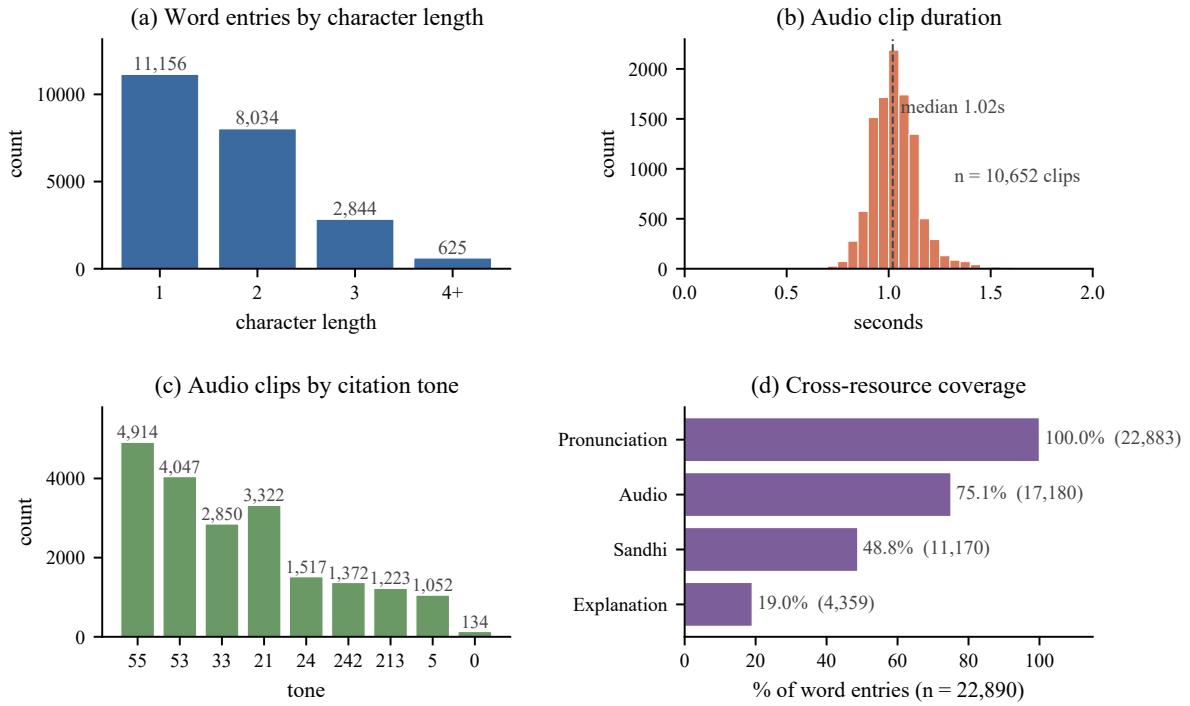

Figure 2: FuzhouBench v1.0 composition. (a) 22,659 published word entries are dominated by mono- and disyllabic forms, with a long tail of three- and four-syllable expressions. (b) The 10,652 audio clips have a median duration of 1.02 s, reflecting their dictionary-style (single word or short phrase) nature rather than narrative speech. (c) Citation tone coverage spans the full Fuzhou 7+5 inventory, with the four most frequent open tones (55, 53, 33, 21) accounting for the majority of clips. (d) Each word entry is covered to a different depth by the other resources: nearly every word has a pronunciation, three quarters have audio, just under half carry an explicit tone-sandhi annotation, and one in five has a written explanation.

torical phonology metadata.

4.2 Audio

The audio portion is the source community's recording project, in which two native speakers (anonymized as Speaker 1 and Speaker 2) recorded dictionary-style readings of words and short phrases. Files are MP3 with mixed sample rates (48 kHz and 44.1 kHz) and mixed channel counts (mono for Speaker 1, stereo for Speaker 2). We document the raw state in Section 5 and apply uniform preprocessing only at task time.

4.3 Quality Control

4.3.1 Inherited annotations

Most of the per-entry annotation in FuzhouBench was produced by the source community over multiple years and is inherited rather than newly produced. Yngping romanization for each pronunciation is provided by community editors. Feng-style sandhi forms (the `sandhi_pron` field) are taken directly from the printed Feng dictionary, with editorial cleanup. Lexical category labels in the explanation table follow standard part-of-speech conventions from the community editor guidelines.

4.3.2 New annotation

We treat the LLM-generated colloquial queries described in Section 6 as a separate evaluation pool, not as gold-standard annotation, and report results on both pools.

4.3.3 Audio quality control

We ran exploratory analysis on all 10,652 audio files. Four files have peak amplitude at or above 0.999 and are flagged as potentially clipped; we keep them but mark them. Median silence ratio is 53%; voice activity detection (VAD) trimming is therefore mandatory before any acoustic task. All file-to-metadata MD5 checks pass.

5 Dataset Statistics

Figure 2 gives a four-panel snapshot: word length, audio duration and tone, and cross-resource coverage. Tables 2–4 report the same numbers in finer detail. The ten tables form a star schema centered on

Resource	Rows	Notes
WordResource	22,890	22,659 published; 11,261 phonology metadata; 20,709 unique surface forms
PronunciationResource	51,505	27,454 sandhi forms; 22,905 primary readings; multi-source[†]
ExplanationResource	5,160	covers 4,359 words; 932 entries include example sentences
FengResource	11,276	7,961 (70.6%) have a surface sandhi that differs from the literal pronunciation
CikLingResource	11,512	historical rhyme-book entries with categorical phonology labels
DFDCharacterResource	9,811	character-level entries with *Bang-uah-ce* romanization
SupplementWordResource	27,871	variant and related forms; type M (main) 19,956, S (synonym) 7,304
RecordingResource	10,659	10,652 unique MD5s, 10,270 unique Yngping forms
UserContribResource	1,059	community contributions, fully reviewed track
UserSubmitResource	121	community contributions, pending review

Table 2: Text resources in FuzhouBench v1.0. All counts are exact and reproduced by the public `recount.py` script. [†]IME 28,908 / Feng 19,861 / community contributions 2,125.

Coverage	Count	%
Words with ≥ 1 Yngping reading	22,883	99.97
Words with Feng sandhi annotation	11,170	48.80
Words with explanation	4,359	19.04
Words with audio (Yngping match)	17,180	75.05

Table 3: Cross-resource coverage among the 22,890 word entries.

Speaker	Files	Hours	Channel
Speaker 1	7,405	2.13	mono
Speaker 2	3,247	0.94	stereo
Total	10,652	3.06	mixed

Table 4: Audio corpus shape. Sample rates: 48 kHz (9,129 files) and 44.1 kHz (1,523 files). Median clip duration is 1.02 s with a 99th percentile of 1.42 s, reflecting the dictionary-style nature of the recordings.

WordResource: Pronunciation, Explanation, Feng, CikLing, DFD, and Supplement are foreign-keyed via `word__id`; Recording links indirectly through the `yngping` field; UserContrib and UserSubmit hold reviewed and pending community submissions (Appendix D).

Pronunciation rows per word. PronunciationResource enumerates each word's literal (citation) form, surface-sandhi form, literary–colloquial variants, and regional variants on separate rows, averaging $51{,}505/22{,}890 \approx 2.25$ rows per word; 22,905 (44.5%) are flagged as the canonical primary reading. These rows are text-only *Yngping readings*; the actual recorded audio lives in the separate RecordingResource. We use "Yngping reading" for the text rows when the distinction from audio matters; the table-name PronunciationResource is kept for fidelity to the source schema.

Coverage and tone inventory. Nearly every entry has at least one Yngping reading (99.97%), supporting full-lexicon G2P training; 48.8% carry a Feng sandhi annotation, yielding ∼11k (literal, sandhi) pairs as sandhi-prediction ground truth; 75% have aligned audio (via Yngping exact match). The audio covers the complete Fuzhou 7+5 tone inventory (seven citation tones used in isolation, plus five sandhi tones used in non-final positions of polysyllabic words; Figure 2c), with the four most frequent open tones (55, 53, 21, 33) accounting for the majority of clips. Table 5 positions FuzhouBench against the few other Sinitic-variety resources: smaller than the Cantonese pretraining corpora, but with a different modality mix.

6 Benchmark Tasks and Baselines

All three tasks use seeded train/test splits with no held-out dev set: every baseline is zero-shot, fixed-prompt, or rule-learned from the train pool, so there are no hyperparameters to tune. Open-model inference runs on a single NVIDIA H100 80GB GPU; closed-model inference uses the OpenAI Batch API.

6.1 Task 1: Grapheme-to-Phoneme

Setup. Given a Chinese character or compound, predict its Yngping. Prior Mandarin G2P work (e.g. Park and Lee, 2020) targets Pinyin polyphone disambiguation; we instead test frontier LLMs few-shot on Yngping, an orthography they were not explicitly trained for. Test sample: 500 words from the 22,883 word–primary-reading pairs, stratified 125 per character-length bucket (1–4), seed 42. Metrics: word-exact, syllable, and tone-only accuracy.

Baselines. We evaluate three open source Qwen3 models (Yang et al., 2025) in a five-shot prompt (Qwen3-1.7B, Qwen3-8B, and Qwen3-30B-A3B; the last is a mixture-of-experts model with 30B total and 3B active parameters per token) and

Resource	Variety	Size	Modalities	Sandhi-annotated
Yue-Benchmark	Cantonese	3.7k MMLU-style items	text	no
YueData	Cantonese	2B tokens	text	no
Hokkien Gezi Opera	Hokkien	3 hours singing	audio	no
CompLex-ZH	Mandarin+Cantonese	1.8k words	text	no
FuzhouBench	Eastern Min	22.9k words + 3.06 h	text + audio + IPA	**yes (11.3k)**

Table 5: FuzhouBench compared with recent Sinitic-variety resources. We are the first to include explicit tone-sandhi annotations and a paired audio-text release for a Sinitic variety other than Cantonese and Hokkien.

System	Word	Syl.	Tone
Qwen3-1.7B	0.00	0.58	6.48
Qwen3-8B	0.40	1.00	16.21
Qwen3-30B-A3B	0.80	0.83	8.56
gpt-5.4-mini	**2.40**	**3.57**	**24.52**

Table 6: G2P baselines on the 500-word stratified test sample (125 per character-length bucket, seed 42). All numbers are percentages. The closed `gpt-5.4-mini` model outperforms the best open Qwen3 model on every metric by 3 to 4x, but absolute scores remain low: under 3% of words are predicted exactly.

System	Word	Syl.	Tone
Identity	7.60	33.19	53.52
Mode-rule (1-tone)	14.40	42.92	67.33
Context-rule (2-tone)	**34.60**	**56.56**	**85.75**
Qwen3-1.7B	21.20	42.48	68.81
Qwen3-8B	23.40	44.92	70.03
Qwen3-30B-A3B	23.80	45.79	71.76
gpt-5.4-mini	27.20	50.48	77.24

Table 7: Sandhi prediction baselines on a 500-entry multi-syllable test sample. All numbers are percentages. LLMs receive a pattern-matched 5-shot prompt: each test item is paired with five demonstrations that share its first non-final (curr, next) tone pattern. The 62-entry symbolic context-rule (the standard descriptive Fuzhou sandhi grammar) leads every LLM by 7–13 points on word-exact accuracy.

one closed source model (`gpt-5.4-mini`[3] (OpenAI, 2025) via the OpenAI Batch API with `reasoning_effort=none`). Each model receives the same five hand-picked Yngping demonstrations, chosen from the training data to cover both monosyllabic and disyllabic forms and both open and checked tones; open models are decoded greedily (`do_sample=False`) with a 40-token cap.

6.2 Task 2: Tone Sandhi Prediction

Setup. Given a multi-syllable word's literal (citation) Yngping, predict the surface sandhi pronunciation. The data are the 11,276 Feng entries that carry both forms; 70.6% have a sandhi form that differs from the citation. We hold out a 500-word multi-syllable test sample (seed 42); the remaining 8,072 form the train pool. Metrics: word-exact, syllable, and tone accuracy (§2 sketches why Fuzhou sandhi is harder than Mandarin or Cantonese sandhi).

Baselines.

- **Identity**: predict sandhi = literal verbatim. Exposes the sanity floor.
- **Mode-rule (1-tone)**: a single-tone-context lookup learned from the train pool. For each literal tone in non-final position, predict the majority sandhi tone (e.g. 53→21, 33→21, 242→53); the final syllable's tone is kept unchanged. Intentionally context-blind; serves as a lower bound on what symbolic knowledge with no context can buy.
- **Context-rule (2-tone)**: a (literal_t_i, literal_t_{i+1}) → sandhi_t_i lookup, also learned from the train pool. This corresponds to the standard descriptive Fuzhou sandhi grammar, which is conditioned on the following syllable's tone. Learning the table from the train pool yields 62 (curr, next) contexts; unseen contexts back off to the 1-tone mode rule.
- **LLM few-shot (pattern-matched)**: Qwen3-1.7B, Qwen3-8B, Qwen3-30B-A3B (open), and `gpt-5.4-mini` (closed, via the OpenAI Batch API with `reasoning_effort=none`). For each test item the five demonstration examples are sampled to share the test item's first non-final (curr, next) tone pattern. This is the fair test of in-context rule application: the model sees the exact mapping it needs to apply, instantiated five different ways.

[3]Parameter count not publicly disclosed; we treat it as a black-box closed comparator.

6.3 Task 3: Definition Retrieval

Setup. Given a colloquial Mandarin query, retrieve the most relevant Fuzhou entry from the 22,883 entries with at least one pronunciation. Evaluation: Hit@1, Hit@5, Hit@10, MRR.

Two evaluation pools. We evaluate on two pools that stress different retrieval regimes:

- *Original-source queries*: 8,880 queries automatically extracted as the first sentence of each entry's explanation.
- *LLM-generated colloquial queries*: 11,225 queries produced by Qwen3-30B-A3B from the explanations, prompted to rewrite each explanation as a natural Mandarin search phrase.

Baselines.

- BM25 (Robertson and Zaragoza, 2009) with character-level and 2-gram tokenization.
- Vector retrieval with `bge-large-zh-v1.5` (Xiao et al., 2023) (326M parameters, 1024 dimensions), a Chinese-specialized BERT (Devlin et al., 2019; Cui et al., 2021) sentence encoder in the Reimers and Gurevych (2019) bi-encoder style; this is the standard dual-encoder dense-retrieval setup (Karpukhin et al., 2020).
- Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) of BM25 and vector, with $K = 60$.
- Two-stage: RRF retrieval followed by a cross-encoder reranker (`bge-reranker-v2-m3` (Chen et al., 2024), 568M parameters), reranking the top 50 to top 10. Figure 4 in Appendix E sketches the full pipeline.

Results are in Table 8.

Results. The contrast between the two pools is the main finding. BM25 is unbeatable on dictionary-original queries because the queries are essentially duplicates of indexed text. On colloquial queries, which is the realistic use case, BM25 loses by a wide margin and a two-stage pipeline is necessary to reach acceptable Hit@10. The reranker is the single largest improvement in the pipeline on both pools: $+0.150$ MRR on original-source queries and $+0.156$ on colloquial queries, with the colloquial post-rerank result (MRR 0.716) closing most of the gap with the best in-domain BM25 number (0.799 on original-source).

Embedding ablation. The choice of vector encoder matters less than its language coverage. Table 9 compares three encoders on the colloquial-query pool: a 118M-parameter multilingual mini-BERT, the 326M Chinese-specialized BGE used as our baseline, and a 1.5B multilingual LLM-based encoder. The two Chinese-aware options (bge-large at 326M and gte-Qwen2 at 1.5B) sit within 0.002 MRR of each other despite a $4.6\times$ parameter gap; the smaller multilingual MiniLM (Reimers and Gurevych, 2020) collapses by 0.43 MRR and is the only configuration in our entire benchmark that loses to BM25 on the same query pool (0.257 vs. BM25 0.372). Comparable patterns of multilingual generality at the cost of per-language depth are documented for XLM-R (Conneau et al., 2020) and other multilingual encoders. Chinese specialization is more important than parameter count here.

7 Analysis

7.1 Sandhi compliance is high but not uniform

Cross-checking the eighteen documented Fuzhou sandhi rules against the audio (F0 contour vs. citation contour for the predicted target tone) shows broad agreement: 11 of 18 rules reach $\geq 77\%$ directional compliance over $n \geq 30$ tokens. Three rules are strongly contradicted (53→33: 12% over 445 tokens; 5→33: 2% over 130; 33→21: 42% over 373); whether this is idiolectal drift, citation-form artifact, or a real grammar change is left to a companion phonetics paper. These three rules cover $\sim$8% of Feng sandhi pairs, giving an upper bound on rule-based exact-match accuracy.

7.2 Why colloquial queries change everything

In retrieval the dominant factor is query distribution: BM25 wins by 21 points on dictionary-original queries and loses by 31 on colloquial ones, while the vector model is roughly invariant. Practitioners who report BM25 as a strong baseline for low-resource dictionary search are typically evaluating on the wrong distribution.

7.3 What the models miss on G2P

The two open- vs. closed-model regimes fail differently. **Qwen3 produces Mandarin Pinyin**: typical Qwen3-30B-A3B predictions like "sang5 han242" for gold *suong55 hang53* ("typhoid") or "fa5 ya242" for gold *buok21 nga53* ("to sprout") show three co-occurring failures—initials and fi-

	Original-source queries (8,880)				LLM colloquial queries (11,225)			
System	H@1	H@5	H@10	MRR	H@1	H@5	H@10	MRR
BM25	0.719	0.903	0.938	0.799	0.282	0.492	0.572	0.372
Vector (bge-large)	0.490	0.727	0.798	0.591	0.575	0.806	0.866	0.684
RRF (BM25 + vector)	0.630	0.856	0.911	0.727	0.435	0.728	0.824	0.560
+ Reranker (v2-m3)	**0.817**	**0.952**	**0.971**	**0.877**	**0.621**	**0.842**	**0.900**	**0.716**

Table 8: Definition retrieval baselines on both evaluation pools. BM25 dominates when queries are extracted from the dictionary itself; vector retrieval becomes competitive once queries are rewritten into colloquial Mandarin. The cross-encoder reranker (`bge-reranker-v2-m3`) lifts MRR on both pools and is the single largest improvement in the pipeline: $+0.150$ MRR on original-source queries ($0.727 \rightarrow 0.877$) and $+0.156$ on LLM colloquial queries ($0.560 \rightarrow 0.716$).

Embedding	Params	Dim	Vec MRR
MiniLM-L12 (multilingual)	118M	384	0.257
bge-large-zh-v1.5 (Chinese-specialized)	326M	1024	**0.684**
gte-Qwen2 (multilingual)	1.5B	1536	0.686

Table 9: Embedding ablation on the 11,225 colloquial query pool, vector-only retrieval before reranking. The Chinese-specialized bge-large sits within 0.002 MRR of a 4.6×-larger multilingual LLM-based encoder, while a smaller multilingual mini-BERT collapses by 0.43 MRR and even loses to BM25 (0.372) on this pool. Bolded row is our default.

nals spelled as Mandarin Pinyin (*sh-, f-, j-* rather than Fuzhou *s-, b-, z-*), tone digits drawn from Mandarin 1–4 or a small fallback set (55/213/242/5), and dropped checked-tone codas. **gpt-5.4-mini sits in between**: some outputs are recognizably Fuzhou-shaped (e.g. "cing55 ga55 mu33" for gold *cing21 nga53 mu33*), but tones are wrong even when the segments are right, and the model occasionally mixes IPA (ŋ) into the output. Across all four LLMs, scaling from 1.7B to 30B-A3B does not help; the closed model lifts syllable accuracy a few times but word-exact stays under 3%. Targeted fine-tuning, RAG with the Qi-Lin rhyme-book tone categories, or hybrid rule-based intervention all remain open.

7.4 LLMs imitate the rule but do not internalize it

Pattern-matched 5-shot lifts every LLM to 21–27% word-exact (best `gpt-5.4-mini` at 27.2%) from 12–15% in a random 5-shot ablation, indicating that LLMs can imitate the local (curr, next) mapping when it is laid out in the prompt but do not carry Fuzhou sandhi as latent knowledge. The remaining 7–13 point gap to the 62-entry rule says they still do not deterministically apply the pattern: they rewrite non-target syllables, default to Mandarin-shaped tones on contexts absent from the demonstrations, and (like the rule) miss Fuzhou's initial-consonant assimilation. Syllable count is where every system collapses: the 2-tone rule reaches 45.97% on disyllabic items but 1.89% on trisyllabic and 0% on 4+, with LLMs no better. The most useful first step for any Fuzhou NLP system is therefore to compose an LLM with a deterministic, context-aware phonology head; the open problem is extending coverage past disyllabic boundaries, not learning sandhi from scratch.

8 Conclusion

We introduce FuzhouBench, the first public NLP benchmark for Eastern Min: 22,890 entries, 51,505 Yngping readings, 11,276 sandhi annotations, and 3.06 hours of audio, with three tasks and complete baselines. Three findings stand out. (1) Yngping G2P is essentially unsolved: Qwen3 1.7B–30B-A3B stay under 1% word-exact and `gpt-5.4-mini` reaches only 2.40%, with both regimes producing Mandarin-Pinyin-shaped output. (2) A 62-entry symbolic context-rule for sandhi beats every LLM by 7–13 word-exact points even with pattern-matched 5-shot prompting; LLMs imitate the local mapping but do not internalize the rule. (3) On colloquial Mandarin queries, BM25 collapses (MRR 0.799→0.372) and a two-stage hybrid + cross-encoder pipeline recovers MRR to 0.716, with the reranker contributing the single largest improvement. Concrete next steps: extend the sandhi rule beyond disyllabic contexts (the 62-entry table drops to 1.9% word-exact on trisyllabic items), compose LLMs with a deterministic phonology head, and grow the audio past two speakers.

Limitations

We list limitations in order of how strongly they constrain follow-up work.

Two speakers only. The audio portion comes from two speakers, with one (Speaker 2) contributing almost exclusively monosyllabic citation forms. Any cross-speaker sandhi or TTS analysis is constrained by this design. We document the speaker imbalance explicitly so that follow-up work can either expand the recording effort or design around it.

Dictionary-style audio. Median clip duration is 1.02 seconds; the recordings are word- or short-phrase-level rather than narrative. Off-the-shelf TTS fine-tuning recipes assume connected speech and will need adaptation. We note this as a research opportunity rather than a workaround.

Limited sub-variety coverage. Fuzhou Eastern Min has substantial sub-variety variation across Lianjiang, Changle, Luoyuan, and other counties. The current release reflects mostly the urban Fuzhou variety as represented in the source dictionary.

Small evaluation set for retrieval gold standard. The native-speaker-written query set is 200 entries. We rely on the larger LLM-generated set for statistical power, with the caveat that LLM-generated queries differ from real user queries in ways that are not yet quantified.

Sandhi baselines model only tone change. Both our 1-tone and 2-tone rule baselines, and the LLM prompts, treat sandhi as a tone-rewriting operation. Fuzhou sandhi also includes initial-consonant assimilation in non-final positions (the citation *s, d, g* that surface as *l, l,* ∅), which the rule baselines do not encode. The reported word-exact ceiling therefore underestimates a fully sandhi-aware predictor; future work that adds the initial-assimilation rules to the table is a clear next step.

Ethics Statement

Data sovereignty. The dataset belongs to the source community. We are releasing an academic-version snapshot. The community retains the right to maintain, update, and re-license the underlying database.

Speaker consent. The two audio contributors (Speaker 1 and Speaker 2) agreed to consent their audio for non-commercial research and educational use of their recordings.

Voice cloning risk. Three hours of audio is enough to fine-tune a voice cloning model that could impersonate the speakers. We mitigate this by (a) releasing the dataset under CC-BY-NC-SA 4.0 with explicit non-commercial restriction, (b) committing to not releasing fine-tuned TTS checkpoints capable of zero-shot cloning, and (c) recommending downstream releases include audio watermarking.

Compensation. The source community's contributors are volunteers from the Fuzhounese language community. We did not pay for the inherited annotations because they predate this project.

Anonymization. Before release we (1) dropped all contributor-identity columns from community-submission tables (`user_name`, `contact`, `reader`, all reviewer columns); (2) mapped the two audio speakers' initials to `Speaker_1` / `Speaker_2` in both metadata and audio directory names; (3) regex-scrubbed note and comment text for the source dictionary's name and contributor handles (e.g., `@QQ`, `@Weibo`); (4) stripped ID3 metadata (which encoded recording-session dates) from every audio file in the released sample. We verified the result with automated leak scans on the bundle.

Potential misuse. We do not anticipate any direct misuse vectors beyond voice cloning, addressed above. Eastern Min is a regional language without political controversy in the corpus content; the lexicon does not include hate speech, slurs, or extremist content.

Ethics review. This work did not undergo a formal ethics review board process, as it relies on secondary analysis of an existing, community-maintained dictionary rather than new data collection. The two volunteer audio contributors provided consent specifically authorizing non-commercial research and educational use. For this academic release, we have applied the aforementioned anonymization protocols and restricted downstream usage to the CC-BY-NC-SA 4.0 license. Consequently, this project qualifies as secondary research using de-identified, community-controlled data, which is exempt from primary-research IRB review.

Responsible NLP checklist. We complete the ACL Responsible NLP checklist at submission time.

References

Peng Bai, Yue Zhou, Meizhen Zheng, Wujin Sun, and Xiaodong Shi. 2023. Improving Chinese pop song and Hokkien gezi opera singing voice synthesis by enhancing local modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829.

Yuen-Ren Chao. 1930. A system of tone letters. *Le maître phonétique*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27th edition. SIL International, Dallas, Texas.

Aizhen Feng and Rong Li. 1998. *Fuzhou Fangyan Cidian [Fuzhou Dialect Dictionary]*. Jiangsu Jiaoyu Chubanshe. In Chinese.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Jiyue Jiang, Alfred Kar Yin Truong, Yanyu Chen, Qinghang Bao, Sheng Wang, Pengan Chen, Jiuming Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2025a. Developing and utilizing a large-scale Cantonese dataset for multi-tasking in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Jiyue Jiang, Alfred Kar Yin Truong, Yanyu Chen, Qinghang Bao, Sheng Wang, Pengan Chen, Jiuming Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2025b. How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Shuheng Liu and Michael Best. 2025. A survey of NLP progress in Sino-Tibetan low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7804–7825.

Sin-En Lu, Bo-Han Lu, Chao-Yi Lu, and Richard Tzong-Han Tsai. 2022. Exploring methods for building dialects-Mandarin code-mixing corpora: A case study in Taiwanese Hokkien. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Robert Samuel Maclay. 1870. *An alphabetic dictionary of the Chinese language in the Foochow dialect*. Methodist Episcopal Mission Press.

OpenAI. 2025. GPT-5 system card.

Kyubyong Park and Seanie Lee. 2020. g2pM: A neural grapheme-to-phoneme conversion package for Mandarin Chinese based on a new open benchmark dataset. *Preprint*, arXiv:2004.03136.

Le Qiu, Shanyue Guo, Tak-Sum Wong, Emmanuele Chersoni, John Lee, and Chu-Ren Huang. 2024. CompLex-ZH: A new dataset for lexical complexity prediction in Mandarin and Cantonese. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Rong Xiang, Ming Liao, and Jing Li. 2024. Cantonese natural language processing in the transformers era: A survey and current challenges. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general Chinese embedding. *Preprint*, arXiv:2309.07597.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yi Yang, Yiming Wang, ZhiQiang Tang, and Jiahong Yuan. 2024. Automated tone transcription and clustering with Tone2Vec. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

A Datasheet

We provide the full datasheet in supplementary materials. Key fields summarized here:

- **Motivation.** First public Eastern Min NLP benchmark; needed because no comparable resource exists.
- **Composition.** Text (TSV) and audio (MP3); see Section 5.
- **Collection.** Inherited from the source community's open dictionary; no scraping.
- **Preprocessing.** Audio resampled to 24 kHz mono with VAD trim at task time; text used as-is.
- **Uses.** Recommended: G2P, sandhi prediction, definition retrieval, TTS research, dialectology. Not recommended: training voice cloning models, commercial use without permission.
- **Distribution.** HuggingFace Datasets and Zenodo, with DOI.
- **Maintenance.** Source community via lead maintainer; semantic versioning.

B Tone Inventory

We use Chao (1930)'s five-level pitch notation: each digit is a pitch level from 1 (lowest) to 5 (highest), and a multi-digit string is the sequence of pitch targets traversed by the syllable. For example, 55 denotes a flat high tone, 53 a high-to-mid fall, and 242 a mid-high-mid contour.

The nine surface contours used in FuzhouBench, with their frequencies in the 10,652-clip audio corpus, are:

- **55** (high level): the most frequent open tone, $n = 4{,}914$.
- **53** (high-to-mid falling): $n = 4{,}047$.
- **21** (low falling): $n = 3{,}322$.
- **33** (mid level): $n = 2{,}850$.
- **24** (mid rising): $n = 1{,}517$.
- **242** (mid-high-mid): $n = 1{,}372$.
- **213** (low falling-rising): $n = 1{,}223$.
- **5** (high checked): a short syllable closing on a glottal stop or velar stop coda (–h, –k); $n = 1{,}052$.
- **0** (neutral): used on unstressed clitics and grammatical particles, with pitch determined by the surrounding context rather than by a fixed contour; $n = 134$, the rarest in the corpus.

Open tones (55, 53, 33, 21, 24, 213, 242) appear on syllables ending in a vowel, nasal, or liquid; the checked tone (5) appears on syllables ending in a stop coda; the neutral tone (0) appears on unstressed syllables. Descriptive grammars vary on the precise category labels (yinping, yangping, etc.) attached to these contours, especially for 21, 24, 213, and 242, and we therefore identify tones by their Chao numerals throughout the paper.

C Mandarin–Fuzhou Lexical Comparison

Section 2 notes that many everyday concepts use entirely different morphemes in Fuzhou and Mandarin, independent of the phonological alternations. Table 10 catalogs 14 representative pairs drawn from native-speaker knowledge and cross-checked against the dictionary. We group them into three categories: single-morpheme substitutions (different character for the same concept), verb–object collocation rewrites (different verb and/or noun, sometimes with a different underlying metaphor such as "eat-smoke" for "smoke cigarettes"), and compound-noun rewrites.

Mandarin	Fuzhou	Yngping	Meaning
Single-morpheme substitutions			
锅	鼎	diang33	pot
筷子	箸	deoy242	chopsticks
屋	厝	cuo213	house
站	徛	kie242	stand
蛋	卵	loong242	egg
Verb–object collocation rewrites			
吃早饭	食早	sieh33 za33	eat breakfast
下车	落车	loh55 jia55	get off (vehicle)
下雨	逷雨	doung53 ngy33	rain (v.)
抽烟	食熏	sieh5 houng55	smoke (cigarette)
喝酒	食酒	sieh33 ziu33	drink (alcohol)
回家	轉厝	duong55 cuo213	go home
洗脸	洗面	se55 meing213	wash face
理发	剃頭	tie55 lau53	get haircut
Compound-noun rewrites			
脸毛巾	面布	ming53 muo213	face towel

Table 10: Mandarin vs. Fuzhou lexical comparison for 14 common everyday concepts. The two languages select different morphemes for the same referent (independent of sandhi or initial-consonant assimilation). The Fuzhou column uses traditional Min orthography because several characters (e.g., 厝, 逷, 徛) have no simplified equivalent; the Mandarin column uses standard Simplified for comparison.

This lexical gap is the direct structural reason BM25 collapses on colloquial-Mandarin queries (Section 6.3): a query phrased in Mandarin will not share surface tokens with a dictionary entry indexed under the Fuzhou form. A two-stage hybrid + cross-encoder retriever partially recovers because the dense embeddings and reranker have learned cross-lingual concept alignment from broad pretraining; BM25 has no such fallback.

D Dataset Schema

The ten resource tables form a star schema centered on WordResource, the canonical word entity. Figure 3 sketches the relations.

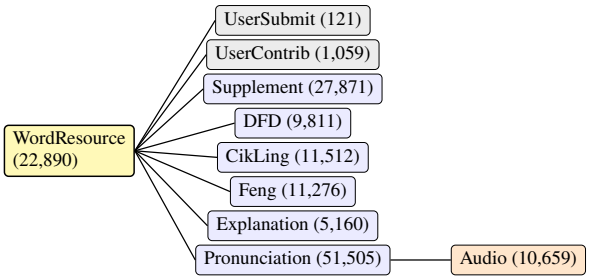

Figure 3: Star-schema view of the ten resource tables (row counts in parentheses; the "Audio" node corresponds to RecordingResource in the source schema, shortened here for layout). Edges leaving WordResource are `word__id` foreign keys. The single Pronunciation → Audio edge is the one indirect link: it matches by exact `yngping` string rather than by ID, because the recording sessions predate the current Pronunciation IDs. UserContrib and UserSubmit are the community-submission queue (reviewed and pending, respectively); validated rows are promoted into the main tables.

Most analyses join Pronunciation onto Word and then optionally onto Feng (for sandhi) or Recording (for audio). The 70.6% non-trivial sandhi rate cited in Section 6.2 is computed on the Pronunciation–Feng join. The 75.1% audio coverage in Table 3 is computed by counting distinct Word entries that have at least one Pronunciation whose `yngping` appears in Recording.

E Task Prompts and Sample Responses

This appendix shows the prompt structure used for the three benchmark tasks and one representative sample for each. We highlight the expected (gold) answer in green and the model's actual response in red. Few-shot demonstrations that originally include Chinese characters are described by English gloss here for typesetting; the verbatim prompts (with the original characters) are in the released code.

E.1 G2P (fixed 5-shot)

System. You convert Fuzhou Eastern Min words written in Chinese characters to their Yngping romanization. Each syllable in Yngping is letters followed by tone digits from $\{55, 53, 33, 21, 24, 242, 213, 5, 0\}$. Multi-syllable outputs are space-separated. Output ONLY the Yngping string, no commentary.

User. Five fixed (character, Yngping) demonstrations followed by the test word; the five demonstrations gloss as "milk" (→ `ngu33 nei242`), "to speak" (→ `gong242`), "Chinese cabbage" (→ `ba5 cai213`), "to do what" (→ `co213 sa242`), and "I/me" (→ `nguai242`).

Sample (`gpt-5.4-mini`). Test word: the Fuzhou word for "flat persimmon."

- Gold: `bieng55 kei242`
- Model: `biang53 si242`

The model uses Mandarin-Pinyin-style segments (*biang*, *si*) where Yngping has *bieng*, *kei*, and gets every tone digit wrong.

E.2 Sandhi Prediction (pattern-matched 5-shot)

System. You are a Fuzhou Eastern Min phonology expert. Given a literal (citation-form) Yngping string for a multi-syllable word, output its surface (sandhi-form) Yngping. Fuzhou tone sandhi mostly changes the tone digits of non-final syllables; segments stay the same. Tone digits from $\{55, 53, 33, 21, 24, 242, 213, 5, 0\}$. Output ONLY the sandhi Yngping string.

User. Five demonstrations whose first non-final syllables share the test item's (curr, next) tone pattern, followed by the test item. For test `dei242 ang213` (pattern `(242, 213)`):

```
Literal: houng242 gau213 gi0
Sandhi: hung53 ngau213 i0
Literal: kooyng242 sau213
Sandhi: keoyng53 nau21
Literal: dang53 zei242 dang213
Sandhi: dang21 zi53 lang213
Literal: ngyong53 mo53 nei242 ga213
Sandhi: ngyong53 mo21 ni53 a213
Literal: miu242 gie213
Sandhi: miu53 ie213
Literal: dei242 ang213
Sandhi:
```

Sample (`gpt-5.4-mini`).

- Gold: `di53 ang213`
- Model: `dei53 ang213`

The model gets the tone change `242`→`53` right (the dominant pattern in the demonstrations), but misses the *dei*→*di* vowel reduction that the demonstrations also exhibit (*houng*→*hung*, *kooyng*→*keoyng*).

E.3 Definition Retrieval

Retrieval is not a prompted-completion task; given a colloquial Mandarin query, the pipeline (BM25 + bge-large vector → RRF → bge-reranker-v2-m3) ranks the 22,883 word entries (Figure 4). We show one representative query.

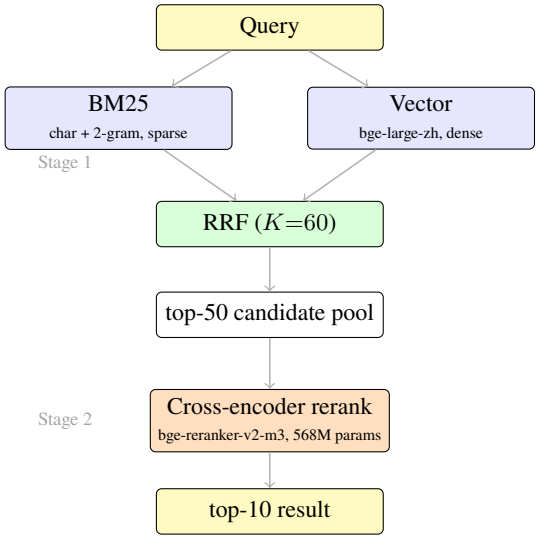

Figure 4: Two-stage retrieval pipeline used for Task 3 Definition Retrieval. **Stage 1 (blue):** BM25 (Robertson and Zaragoza, 2009) and the bge-large-zh (Xiao et al., 2023) vector retriever score the 22,883 word entries in parallel; their ranked lists are fused by Reciprocal Rank Fusion (Cormack et al., 2009) with $K = 60$ to form a single 50-candidate pool. **Stage 2 (orange):** the bge-reranker-v2-m3 (Chen et al., 2024) cross-encoder rescores each (query, candidate) pair jointly and returns the top 10. Section 6.3 reports both pre- and post-rerank metrics.

Query (LLM-generated colloquial Mandarin). "a person who runs around all over the place" (到处乱跑的人), generated by Qwen3-30B-A3B from the explanation of a Fuzhou-specific dialect word.

Top-1 result.

- Gold entry: a dialect word "野蹰" literally translatable as "wild walking," Yngping `ia55 bie213`, glossed in the source dictionary as "to wander around without restraint."

- Pipeline top-1 (after rerank): *same entry*; recovered at rank 1
- BM25 only: ranked the gold entry outside the top 10 because the query shares no character with the entry's Mandarin gloss.

This is a typical pattern: BM25 misses on colloquial queries and the dense reranker brings the dialect-specific entry to the top.

F Reproduction Notes

The data statistics in Section 5 are produced by a single script (`recount.py`) that reads the released TSVs and audio EDA summary and writes a JSON snapshot. Re-running the script on a fresh download should reproduce the numbers exactly, modulo any subsequent version bumps to the source database.