

Taking part

Everything a team meets, from the public board to a published result. Every screen here is the live platform, and every caption was checked against it as this deck was built.

What this measures

Production Cantonese voice systems, scored on real Hong Kong insurance calls rather than read speech. Four pillars, one score, and every number published with the uncertainty around it. There is no public sign-up — accounts are issued by the organisers, because an account is the only thing between a stranger and real customers’ recorded voices.

POINT AT

The four pillar cards. Each links to exactly how that pillar is scored.

✓ Verified live: this screen said “Cantonese Voice Benchmark” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Sign in

CANTONESE VOICE BENCHMARK 2026.1 · CHALLENGE ROUND 21

Cantones voice,
measured on real calls.

Production voice systems evaluated on Hong Kong insurance contact-centre audio across four pillars — hearing, understanding, speaking, and conversation.

Enter the benchmark

How we measure

SIX TONES · THE WORDS OF A CLAIM

單

daan1

policy

326×

保

bou2

to insure

755×

費

fa13

premium

218×

賠

pui4

to pay out

55×

買

maa15

to buy

67×

額

ngaak6

sum insured

39×

Six tones, six words — and every one of them is said in the calls this benchmark is scored on. Play one: the line is its pitch, and getting that shape wrong is what swaps one word for another.

Window openClips drawn 30Entrants 0Updated 02 Sept, 01:56 HKT

No results published yet

The first scored window has not closed. Reference baselines are shown below so the scale is meaningful the moment real entries land.

RANKTEAMCORE

ASR
CER

LLM
JUDGE

TTS
MOS

E2E
NATURAL

TTFB

REFERENCE BASELINES

1

How to read the board

Every published figure carries a confidence interval. Entries whose intervals overlap are reported as tied rather than ordered — at these clip counts a fraction of a percent is noise, and a board that ranks noise is a board that misleads.

POINT AT

Core against Full, and the difficulty filter: the same teams, three honest views.

✓ Verified live: this screen said “Leaderboard” and “confidence interval” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Sign in

CANTONESE VOICE BENCHMARK 2026.1 · CHALLENGE ROUND 21

Leaderboard

Every published number carries a bootstrap confidence interval. Entries that overlap are reported as tied rather than ordered — with 30 clips a fraction of a percent is noise.

WindowChallenge round 21

Closes09 Sept, 01:38 HKT

Entrants0

Updated02 Sept, 01:56 HKT

RANKING

Core ASR + LLM + TTSFull all pillars

Open the cardDownload PNG

DIFFICULTY

AllEasyNormalHard

No results published yet

The first scored window has not closed. Reference baselines are shown below so the scale is meaningful the moment real entries land.

RANK	TEAM	CORE	ASR CER	LLM JUDGE	TTS MOS	E2E NATURAL	TTFB
REFERENCE BASELINES							
Indicative figures for scale. Not measured on this corpus — no entrant is ranked against them.							
A	Open-source pipeline	76.2	11.1%	3.81 / 5	3.44 / 5	3.29 / 5	2.11s p95 3.88
B	Proprietary voice LLM	82.7	8.6%	4.14 / 5	3.62 / 5	4.01 / 5	0.88s p95 1.52

How we measurePermanent link to this edition

Snapshot 02 Sept, 01:56 HKT · scoring v1.2

Recognition — 20%

Can you hear? Six tones, constant English code-switching, insurance vocabulary, and callers who are frightened or angry. Required of every entrant.

POINT AT

The weight and whether the pillar gates the result, stated before anything else.

✓ Verified live: this screen said “Can you hear?” and “20% OF THE SCORE” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Sign in

ASR · 20% OF THE SCORE · EVERY ENTRANT

Can you hear?

Transcribing Hong Kong Cantonese: six tones, constant English code-switching, insurance vocabulary, and callers who are frightened or angry.

Six tones, six words.

Cantonese carries six lexical tones. A tone error does not degrade a word, it replaces it. Every tile below is a word from the calls this benchmark scores — one per tone, with how often it is said in them. Say one on the wrong contour and you have said a different word, which is what the pairs further down price: 買 maa15 / 賣 maa16. Play them.

 單 daan1 policy 326x 55 · HIGH LEVEL	 保 bou2 to insure 755x 25 · HIGH RISING	 費 fa13 premium 218x 33 · MID LEVEL	 賠 pu14 to pay out 55x 21 · LOW FALLING	 買 maa15 to buy 67x 23 · LOW RISING	 額 ngaak6 sum insured 39x 22 · LOW LEVEL
---	--	--	--	--	---

Pitch demonstration, synthesised from the Chao values shown. Not a recording of speech — the point is the shape, not the voice.

WHERE IT COSTS MONEY

Understanding — 25%

Can you answer? A system that transcribes a call perfectly and loses the policy number has not understood it. Judged on whether the answer is grounded in what was said.

POINT AT

What the judge scores, per answer.

✓ Verified live: this screen said “25% OF THE SCORE” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Sign in

LLM · 25% OF THE SCORE · EVERY ENTRANT

Can you understand?

Given the call, does the system grasp what the caller needs, ground its answer in the actual policy, and refuse to invent coverage that does not exist?

Both of these are fluent.

One of them costs the insurer a claim. Hallucinated coverage does not look like an error in a transcript — it looks like good service, which is exactly why it has to be scored separately from whether the words were transcribed correctly.

THE CALLER ASKS

我做physiotherapy，一個療程十次，一次八百蚊，我份plan cover唔cover到？
I need physiotherapy — ten sessions at \$800 each. Does my plan cover it?

GROUNDDED

你張保單有物理治療嘅保障，不過我喺度睇唔到每次嘅上限同埋年度限額，要幫你查返份保單條款先可以講到準確數字。要唔要我而家幫你開個查詢，跟進返俾你？

Says what the policy shows, names what it cannot see, and offers the next step. It answers less than the caller wanted, which is the correct amount.

INVENTED

有嘅，你份plan完全cover到物理治療，每次八百蚊都喺保障範圍之內，一年最多可以做二十次，你直接去做就得，唔使預先approve。

Every marked clause is a fact the policy documents do not contain: the coverage, the per-session limit, the annual cap, and the claim that no pre-approval is needed. The caller books ten sessions on the strength of it.

5 OF 15

Speech — 30%

Can you speak? The least-solved piece of the stack. Delivery is 60% and tone survival 40% — a flat contour turns "buy" into "sell", which in an insurance call is not a cosmetic error.

POINT AT

Every entrant: speech is one of the three pillars in the Core Score.

✓ Verified live: this screen said “30% OF THE SCORE” when this slide was made.

Cantonese Voice Benchmark

粵語語音基準

Leaderboard

Pillars

Methodology

Report

Approaches

Dataset

Enter

EN / 繁中

Sign in

TTS · 30% OF THE SCORE · EVERY ENTRANT

Can you speak?

The least-solved piece of the stack. No dominant Cantonese voice exists, and the reasons are structural rather than a matter of more training data.

Six contours, produced.

Recognition has to tell 6 tones apart. Synthesis has to make them distinguishable in the first place, and the ways that fails do not show up as a bad word — they show up as a voice a Hong Kong listener does not trust. These are the four we hear most.

Tone flattening 婦 fu5 → 父 fu6	Low rising collapses into low level under fast speech or at the end of a phrase. Wife becomes father. The sentence stays fluent and the person changes, and on a call that names a beneficiary there is no context to repair it from.
Reading written Chinese aloud 沒有 → 冇	Output is grammatical Standard Written Chinese pronounced in Cantonese, rather than spoken Cantonese. Every sentence is correct and nobody talks like that. It is the single most common way a synthesised Hong Kong voice announces itself as one.
Code-switch prosody 你個 claim 已經 approve 咗	The English words are produced with English stress and timing, so the sentence audibly changes speaker mid-phrase.

6 OF 15

Conversation — 25%

Can you converse? Audio in, audio out, on the clock. Latency is rewarded here and nowhere else, because this is the pillar where a pause is the product.

POINT AT

Time to first sound after the caller stops.

✓ Verified live: this screen said “25% OF THE SCORE” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Sign in

E2E · 25% OF THE SCORE · OPTIONAL

Can you converse?

Audio in, audio out, on the clock. Not whether the parts work, but whether the whole thing holds a conversation someone would stay on.

The gap is the product.

Everyone agrees that under 300ms feels instant, and nobody knows what that is like. Play a row: the caller finishes, the wait happens at full speed, then the system starts. Same sentence every time — only the silence changes.

Under 300ms 240ms

scores 100/100

Inside the gap a person leaves between turns. Reads as listening.

Around a second 1000ms

scores 59/100

Noticeable. The caller starts wondering whether it heard them.

Over two seconds 2300ms

scores 0/100

The caller says "hello?" and now both of you are talking.

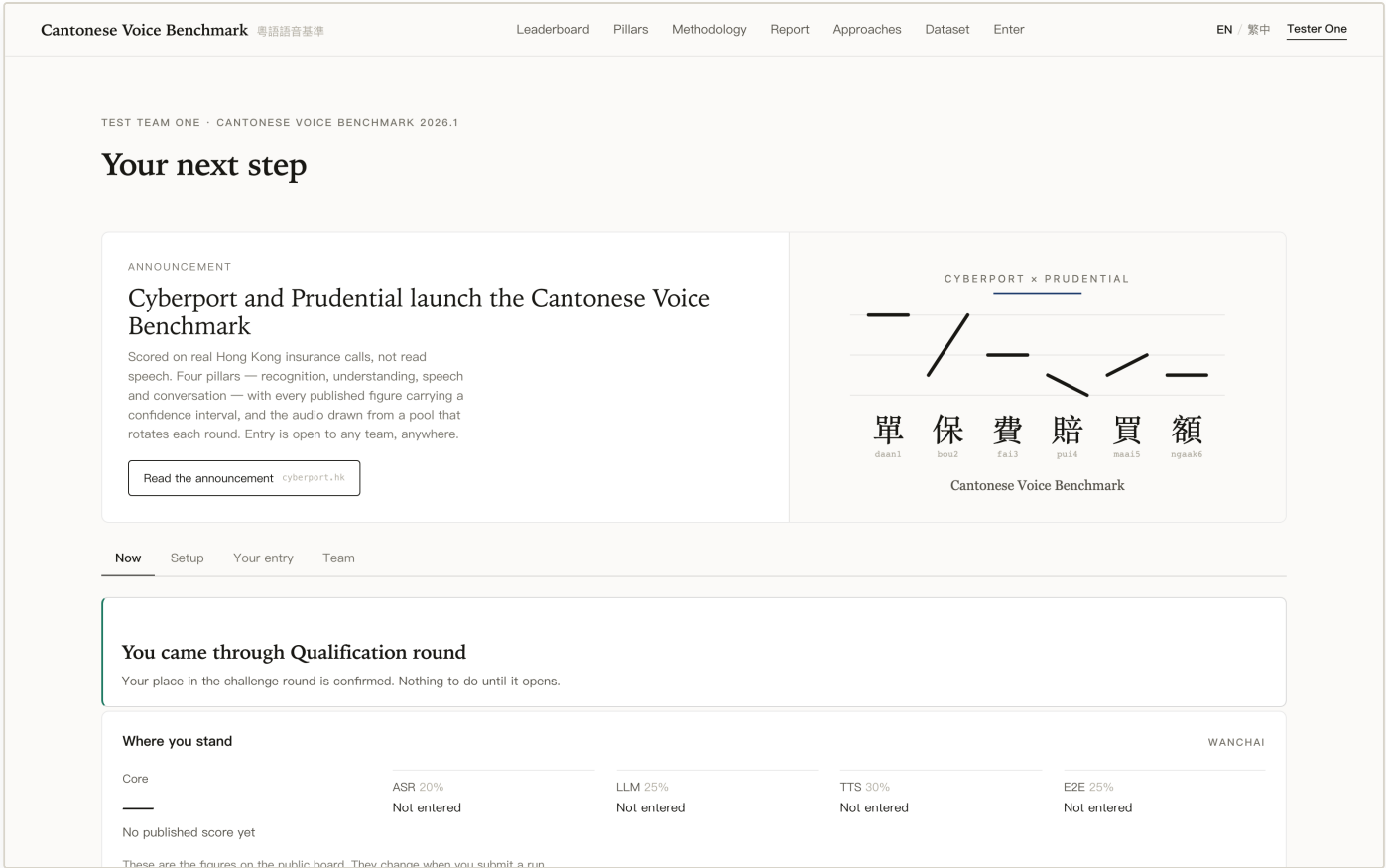
Your portal

One screen answering two questions: what do I do next, and where do I stand. A team that has to work out its own next step is a team that files a support ticket.

POINT AT

Your next step. It moves as you progress and is never blank.

✓ Verified live: this screen said “Your next step” and “Where you stand” when this slide was made.



Registering your system

You run a WebSocket server; we connect to it. Register the URL, then run the connection test — three real clips, about ten seconds, and a readable diagnosis if anything fails.

POINT AT

The starter kit: a working server in Python and Node, so the first green tick takes minutes rather than an afternoon.

✓ Verified live: this screen said “You run a WebSocket server; we connect to it.” and “Start from a server that already runs” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Tester One

TEST TEAM ONE

Endpoints

You run a WebSocket server; we connect to it. Register the URL, then run the connection test — three real clips, about ten seconds, and a readable diagnosis if anything fails.

NowSetupYour entryTeamYour endpoints

Start from a server that already runs

A WebSocket server in Python and Node that speaks our protocol and answers every clip with a fixed line. Run it, register the URL, and the connection test goes green in minutes — then replace one function with your recogniser. Proving the transport before you touch your model is also how you tell our problems from yours later.

Download the starter kit

Try all four pillars against ours

Points recognition, understanding, speech and conversation at the endpoint we run, so you can see every pillar work before you have written a line. It answers all four on one socket. Replace each with your own as you build them.

Use the reference endpoint for all four

Endpoints that fail on purpose

The same service on a different path. Register one, run a practice set, and see exactly what the platform says about that failure — before it happens to you in a scored round. Practice runs cost no quota.

Working reference wss://cvc-reference-entrant.onrender.com	Answers every pillar. Three green checks, a scoreable run, and a poor score — it is a demonstration of the protocol, not a competitor.
Answers too late wss://cvc-reference-entrant.onrender.com/slow	Replies just after the deadline. Shows how a late answer scores: as an empty transcript, with no partial credit and no retry.
Goes quiet wss://cvc-reference-entrant.onrender.com/silent	Accepts the audio and never answers. The worst answer available to you: it costs the full wait and explains nothing.
Answers in Mandarin wss://cvc-reference-entrant.onrender.com/mandarin	Fluent, correct, and the wrong language. Shows the register penalty — the failure a character error rate alone cannot show you.

Practising for free

Eight recordings with their correct transcripts, laid out one line per speaking turn. Click a timestamp to jump there, paste your own attempt, and see the difference character by character. The recordings are real customer calls, so they never leave the platform and never download publicly — and none of this costs a scored run.

POINT AT

Practice is unlimited. Only a scored run spends quota.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Tester One

TEST TEAM ONE · CANTONESE VOICE BENCHMARK 2026.1

Your next step

ANNOUNCEMENT

Cyberport and Prudential launch the Cantonese Voice Benchmark

Scored on real Hong Kong insurance calls, not read speech. Four pillars — recognition, understanding, speech and conversation — with every published figure carrying a confidence interval, and the audio drawn from a pool that rotates each round. Entry is open to any team, anywhere.

Read the announcementcyberport.hk

CYBERPORT x PRUDENTIAL



單保費賠買額

daan1bou2fa1.3pui4naa15ngaak6

Cantonese Voice Benchmark

NowSetupYour entryTeam

Endpoints

Manage

ASR

E2E

LLM

TTS

ok

ok

ok

ok

Your team

Invite colleagues and set what each can do.
One account per person rather than a shared login, so the audit log names who did what.

POINT AT

Roles: who can start a scored run, and who can only look.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Tester One

TEST TEAM ONE

Your team

Everyone here signs in as themselves. Nobody should be sharing a password to watch a leaderboard, and an observer seat costs nothing.

NowSetupYour entryTeamPeople

Names

CODENAME

 Everything published about this team appears as **Wanchai**. Your team name is not published. A codename is the default and stays until you opt in. [Change this on your entry →](#)

Your name

Test Team One

How colleagues see you here. Never published.

Save

Team name

Test Team One

Used on the board only if you choose to be named. Change it any time before then.

Save

PERSON	ROLE	STATE
Tester One tester1@cvc-benchmark.pubrio.com	Owner ▼ Save	ACTIVE
Owner	Everything, including adding people and withdrawing the entry.	
Engineer	Endpoints, runs, the sample pack, the system description. Cannot add people or withdraw.	
Observer	Read-only. Sees results and run history, touches nothing. For an exec sponsor	

Your entry

What you are entering, and which pillars. It is approved the moment you send it. Recognition, understanding and speech make the Core Score, and a rank needs all three. A result is not published until you have also written a description of your system — a leaderboard of anonymous numbers teaches nobody anything.

POINT AT

Approved on sending. Change your answers at any time; it stays approved.

Re-captured 18 Sep 2026: an application is now approved when it is sent.

Cantonese Voice Benchmark 粵語語音基準
Leaderboard
Pillars
Methodology
Report
Approaches
Dataset
Enter
EN / 繁中
Lorewell Engineer

LOREWELL AI

Your application

What you have built and which pillars you are entering. It is approved the moment you send it: the organisers issued your account, and that was the vetting.

Now
Setup
Your entry
Team
Your application

Approved
07 JUL, 16:13 HKT

You are in, for the pillars you entered. Register an endpoint and the portal will walk you through the rest.

Working system, clear description of the stack. Approved for the pillars entered.

Change your answers
It stays approved while you change it. Add a pillar whenever you like; one you already have a scored run in stays.

What have you built?

A production Cantonese speech stack, described in full in the system description.

The organisers read this to know what is behind your entry, and may write to you about it. A few sentences of specifics beats a paragraph of positioning.

Which pillars do you want to enter?

☒ **Recognition** — Can you hear?
20% of the score

☐ **Understanding** — Can you understand?
20% of the score

Your score, and what is behind it

A finished run reports the error rate with the interval around it, the response times, and what the error was actually made of — mishearing, writing Simplified, and answering in Mandarin are three different problems with three different fixes.

POINT AT

The breakdown under the headline number. That is the part you can act on.

✓ Verified live: this screen said “CER” and “What the error is made of” when this slide was made.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中QA tester

CHALLENGE ROUND 21 · SCORED · ASR · LIVE

Run scored

NowSetupYour entryTeamASR run

Result

State	scored
Items	30
Quota consumed	1
CER	98.8% (98.7%—99.0%)
Mix Error Rate	98.7%
Response time (median)	0.01s
Response time (95th)	0.01s
Finished	02 Sept, 01:56 HKT

What the error is made of

Not explained by script or register	98.8%
Cost of writing Simplified	0.0%
Cantonese kept as Cantonese	100.0%
Right words, wrong place	0.4%

Your CER above is unchanged. These say what produced it, because mishearing, writing Simplified, and writing Mandarin are three different problems with three different fixes.

Failures by class

Every item completed cleanly.

Submit this run?

Nobody can see this number but you. A scored run counts towards your published result only once you submit it.

Submitting adds these clips to your standing. It does not replace what you have already

Submitting is your decision

Nobody can see the number but you until you submit it. Submitting ADDS these clips to your standing rather than replacing it — every committed clip is scored together, so more runs narrow the interval instead of moving the number around.

POINT AT

One-way. There is nothing to gain by cherry-picking a lucky run.

✓ Verified live: this screen said “Submit this run” and “It does not replace what you have already submitted” when this slide was made.

Cantonese Voice Benchmark粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中QA tester

CHALLENGE ROUND 21 · SCORED · ASR · LIVE

Run scored

NowSetupYour entryTeamASR run

Result

State	scored
Items	30
Quota consumed	1
CER	98.8% (98.7%—99.0%)
Mix Error Rate	98.7%
Response time (median)	0.01s
Response time (95th)	0.01s
Finished	02 Sept, 01:56 HKT

What the error is made of

Not explained by script or register	98.8%
Cost of writing Simplified	0.0%
Cantonese kept as Cantonese	100.0%
Right words, wrong place	0.4%

Your CER above is unchanged. These say what produced it, because mishearing, writing Simplified, and writing Mandarin are three different problems with three different fixes.

Failures by class

Every item completed cleanly.

Submit this run?

Nobody can see this number but you. A scored run counts towards your published result only once you submit it.

Submitting adds these clips to your standing. It does not replace what you have already

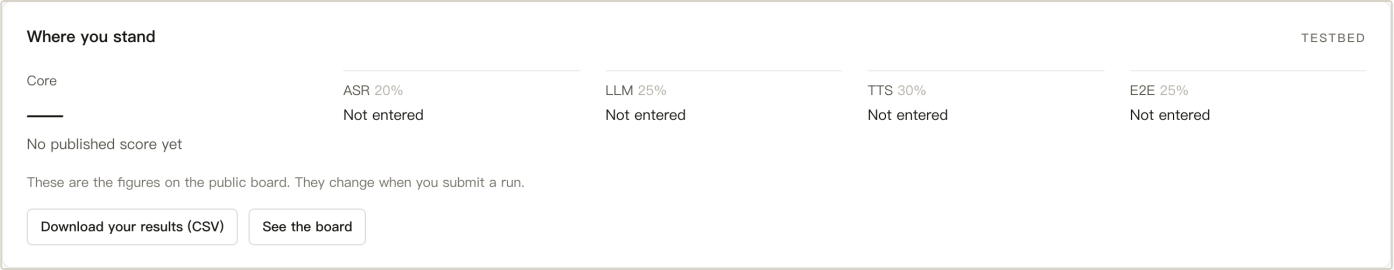
Where you stand

Your Core Score and each pillar, exactly as the public board has them. A pillar you have not submitted reads "not entered" rather than zero, and the figures move only when you submit.

POINT AT

These are the published figures, and your results download as CSV.

✓ Verified live: this screen said “Where you stand” and “These are the figures on the public board” when this slide was made.



Asking us something

Access requests answered in one business day, support the same, security reports the same day if exploitable, and disputes answered with the per-item record behind the number rather than an assurance.

POINT AT

A dispute goes to a queue with a stated turnaround, not an inbox.

Cantonese Voice Benchmark 粵語語音基準

LeaderboardPillarsMethodologyReportApproachesDatasetEnter

EN / 繁中Sign in

SUPPORT

When something is wrong.

Three routes, and what each is for. We would rather you used the fastest one than the politest one.

Platform status

Challenge round 21 is open until 09 Sept, 01:38 HKT. Scored runs are being accepted.

Where to go

A run went wrong	Open the run page first — every clip carries a verdict and an owner, and a failure we caused is already marked as ours and has not consumed your quota. If the record does not match what you saw, open a dispute from that page. Disputes are answered with the per-item evidence behind the number.
Something is broken	If you have an account, raise it from your portal — Team, “Raise something with the organisers”. That goes into the queue an organiser reads, it is counted until somebody answers it, and the reply appears on that same page. Email challenge@pubrio.com only if you cannot sign in: mail is not tracked here, which is why it is the second choice rather than the first. We answer within one business day, and that is a published commitment rather than an aspiration.
A security issue	security@pubrio.com . If it is exploitable, say so in the subject line and we will look the same day. We will not take action against anyone reporting in good faith.