

DexDictate Strict Experiment Matrix (Offline-Only)

Version: 2026-02-19

Scope: Local, private, offline speech-to-text only. No cloud calls.

Goal: Recover and exceed prior speed + accuracy while preserving offline behavior.

Rules (Do Not Violate)

- Keep all tests offline.
- No network calls during dictation runs.
- Test each configuration with identical prompts and environment.
- Record raw timings and transcripts for every run.

Measurement Definitions

- `T_release_to_submit_ms`: Trigger release to `Submitting ... to Whisper` in log.
- `T_submit_to_output_ms`: `Submitting ...` to `Whisper output` in log.
- `T_total_ms`: Trigger release to final pasted text visible in target app.
- `WER_est`: Manual word error rate estimate against reference text.
- `CommandErrorRate`: Wrong command actions / command utterances.
- `PunctErrorRate`: Punctuation errors / punctuation utterances.

Test Environment Lock

Run this before each series:

1. Close all heavy apps (browser tabs, Xcode indexing, video apps).
2. Use same microphone and distance (10-15 cm) for all runs.
3. Keep room noise level constant.
4. Reboot once before Series 1 and Series 5.
5. Use identical target app for paste validation (e.g., TextEdit plain text).

Test Corpus (Use Exactly)

A. General Dictation (accuracy baseline)

1. `DexDictate should transcribe this sentence exactly once.`
2. `Please book a meeting for Tuesday at 3 PM with Dexter.`
3. `I need the quarterly revenue report by end of day.`
4. `The architecture uses event taps and local inference only.`
5. `Do not send data to any external service.`

B. Punctuation / Spacing

1. `This is sentence one period`

2. `This is sentence two period`
3. `Add a comma after this phrase comma then continue.`
4. `Question mark test question mark`
5. `End this line with a period.`

C. Commands / Formatting

1. `new line this should be on a new line`
2. `next line this should also break line`
3. `scratch that`
4. `this line all caps`
5. `hello world scratch that`

D. Hard Words / Names

1. `Kubernetes Istio Prometheus Grafana`
2. `PostgreSQL Redis SQLite Cassandra`
3. `Dexter WestKitty DexDictate`
4. `ane transformer core ml encoder`
5. `whisper cpp swift whisper`

E. Noise Robustness

1. Read A1-A5 with low fan noise.
2. Read A1-A5 with medium keyboard noise.

Series of Instructions

Series 0: Baseline Capture (Current Config)

1. Start clean log capture.
2. Run corpus A-E, 3 repetitions each utterance.
3. Record timings and transcript outputs.
4. Compute means and p95 for `T_submit_to_output_ms` and `T_total_ms`.

Exit criteria:

- Full dataset complete with no missing rows.

Series 1: Accuracy-First Model Sweep

Configs:

- S1A: `tiny.en` (current reference)
- S1B: `base.en`
- S1C: `small.en` (if local latency still acceptable)

For each config:

1. Run corpus A-D, 3 repetitions each.
2. Compute `WER_est`, `PunctErrorRate`, `CommandErrorRate`.
3. Keep identical decoder settings across all three.

Decision rule:

- Promote model if `WER_est` improves by $\geq 20\%$ with `T_total_ms` regression $\leq 35\%$.

Series 2: Decoder Profile Sweep (Per Selected Model)

Profiles:

- S2A Accuracy: `speed_up=false`, fallback enabled, no hard token cap.
- S2B Balanced: dynamic speed profile by utterance length.
- S2C Speed: `speed_up=true`, reduced fallback.

Steps:

1. Run corpus A-D, 3 repetitions each profile.
2. Compare speed/accuracy tradeoff.
3. Reject any profile with command failures $> 2\%$.

Decision rule:

- Pick profile with best weighted score:

```
`Score = 0.55*Accuracy + 0.30*Latency + 0.15*Punctuation`
```

Series 3: Silence/VAD Threshold Sweep

Threshold sets:

- S3A Conservative trim
- S3B Medium trim
- S3C Aggressive trim

Steps:

1. Run corpus A, D, E (focus on start/end clipping risk).
2. Track clipped-first-word incidents and clipped-last-word incidents.

Decision rule:

- Choose fastest threshold set with clipping incidents $\leq 1\%$.

Series 4: Space-After-Period Validation

Methods to compare:

- S4A Trailing-space in pasted text.
- S4B Post-paste synthetic `Space` key event.
- S4C Leading-space on next utterance when previous ended with `.`.

Steps:

1. Run punctuation corpus B, 5 repetitions each method.
2. Validate in TextEdit, Notes, Slack, and browser text area.
3. Count `SpaceMissingAfterPeriod` per method.

Decision rule:

- Choose method with lowest missing-space rate across all target apps.
- If tie, pick method with least side effects in command utterances.

Series 5: End-to-End Soak (Production Candidate)

1. Run selected config for 30 minutes mixed dictation.
2. Include commands every 2-3 utterances.
3. Include at least 20 punctuation endings with period.

Exit criteria:

- No stuck states.
- No transcript drop events.
- `T_total_ms` p95 within target.

Series 6: Regression Gate

Gate must pass before acceptance:

1. `WER_est` \leq baseline - 20%.
2. `T_total_ms` p95 \leq baseline + 10%.
3. `SpaceMissingAfterPeriod` \leq 1/100 utterances.
4. `CommandErrorRate` \leq 2%.
5. No online/network behavior introduced.

Target Thresholds (Strict)

- `T_submit_to_output_ms` mean: \leq 1400 ms (short utterances)
- `T_total_ms` p95: \leq 2200 ms
- `WER_est`: \leq 8% on corpus A+D combined
- `PunctErrorRate`: \leq 5%
- `SpaceMissingAfterPeriod`: \leq 1%

Data Sheet Template (CSV)

Use columns:

- `series,config,utterance_id,repitition,reference,hypothesis,t_release_to_submit_ms,t_submit_to_output_ms,t_total_ms,word_errors,command_error,punct_error,space_after_period_missing,notes`

Test Query Follow (Execution Order)

1. Run `Series 0` first, no tuning changes.
2. Run `Series 1` and lock model choice.
3. Run `Series 2` and lock decoder profile.
4. Run `Series 3` and lock trimming thresholds.
5. Run `Series 4` and lock spacing method.
6. Run `Series 5` soak.
7. Run `Series 6` regression gate.
8. Only then approve as production candidate.

Reporting Format

Produce a final report with:

1. Winning configuration (model + decoder + trim + spacing method).
2. Baseline vs final metrics table.
3. Failures observed and mitigations.
4. Residual risks and next experiment candidates.