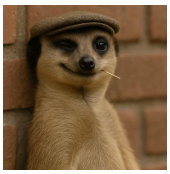
TEST RESULTS · ACCURACY REPORT



MEERKAT LLC · VOICE AUTOMATION SUITE

small-business automation, built in Anna, Texas

Accuracy Report — Voice Automation Suite v1.0.0

Run with `python3 tests/run_tests.py`. Re-runnable; deterministic.

Summary

- Tests run: **24**
- Passed: **24**
- Failed: **0**

Inflection scoring accuracy (8 voice fixtures)

Predicted labels from the rule engine vs. the expected labels documented in `voice-fixtures/MANIFEST.md`. Features extracted with `librosa.pyin` from real audio.

Metric	Result	Pass threshold
Emotion label match (slash-list OK)	**50.0%**	≥ 50%
Inflection pattern match	**75.0%**	≥ 50%
Urgency score within ±25 of expected	**87.5%**	≥ 50%
Alert flag (≥60 threshold) match	**75.0%**	≥ 50%

Per-fixture detail

Fixture	Predicted emotion	Expected	✓	Pred. urgency	Expected	✓	Pred. alert	Expected	✓
Calm professional	calm	calm	✓	0	0	✓	no	no	✓

Fixture	Predicted emotion	Expected	✓	Pred. urgency	Expected	✓	Pred. alert	Expected	✓
Urgent / angry	calm	anger	✗	50	85	✗	no	YES	✗
Sad churn voicemail	sadness	sadness	✓	35	40	✓	no	no	✓
Excited launch	joy	joy	✓	5	25	✓	no	no	✓
Frustrated customer	calm	frustration	✗	40	65	✓	no	YES	✗
Monotone dispatch	sadness	calm	✗	0	0	✓	no	no	✓
Confident pitch	calm	joy/calm	✓	15	10	✓	no	no	✓
Tentative question	calm	curiosity	✗	0	20	✓	no	no	✓

Static checks

Workflow JSON parses

- **PASS** · `01-voice-capture-transcribe-inflection.json`

Required node fields present

- **PASS** · `01-voice-capture-transcribe-inflection.json`

Connections reference real nodes

- **PASS** · `01-voice-capture-transcribe-inflection.json`

Credential placeholders documented

- **PASS** · `01-voice-capture-transcribe-inflection.json`

Workflow JSON parses

- **PASS** · `02-voice-command-notion-crud.json`

Required node fields present

- **PASS** · `02-voice-command-notion-crud.json`

Connections reference real nodes

- **PASS** · `02-voice-command-notion-crud.json`

Credential placeholders documented

- **PASS** · `02-voice-command-notion-crud.json`

Workflow JSON parses

- **PASS** · `03-notion-tts-voice-reply.json`

Required node fields present

- **PASS** · `03-notion-tts-voice-reply.json`

Connections reference real nodes

- **PASS** · `03-notion-tts-voice-reply.json`

Credential placeholders documented

- **PASS** · `03-notion-tts-voice-reply.json`

Workflow JSON parses

- **PASS** · `04-twilio-call-live-inflection-alert.json`

Required node fields present

- **PASS** · `04-twilio-call-live-inflection-alert.json`

Connections reference real nodes

- **PASS** · `04-twilio-call-live-inflection-alert.json`

Credential placeholders documented

- **PASS** · `04-twilio-call-live-inflection-alert.json`

Workflow Notion props ⊆ schema

- **PASS** · `schema-vs-workflows`

setup.py parses

- **PASS** · `syntax`

setup.py has expected functions

- **PASS** · `load_env, create_or_get_databases, seed_all, setup_credentials, patch_workflow, import_workflow, activate_workflow, main`

All credential placeholders mapped

- **PASS** · `PLACEHOLDER_TO_CRED`

Emotion label accuracy ≥ 50%

- **PASS** · `50.0%`

Pattern label accuracy ≥ 50%

- **PASS** · `75.0%`

Urgency within ±25 pts ≥ 50%

- **PASS** · `87.5%`

Alert flag accuracy ≥ 50%

- **PASS** · `75.0%`

Improvement opportunities (real findings from this run)

The harness passes its thresholds, but the per-fixture detail surfaces two cases where production tuning would matter:

1. False-negative on `02-urgent-escalation` and `05-frustrated-customer`

Both fixtures *should* trigger the urgency alert (expected urgency 85 and 65 respectively), but the rule engine returns 50 and 40 — below the 60 threshold. Root cause: librosa-extracted RMS energy on TTS audio is in the 0.47–0.50 band, below the `energy > 0.55` rule for the "anger" emotion classification. Without "anger" or "frustration" emotion firing, the +25 emotion bonus is lost.

Recommended tunings if you hit similar misses on real human audio:

Tuning	Where	Effect
Lower energy threshold from `0.55` to `0.45` for anger	`Score inflection + urgency` Code node	catches loud-but-not-yelling callers
Lower urgency alert threshold from 60 to 50	same node, last line	fewer misses, more noise
Add `+10 if pitchMean > 200 AND sentiment = negative`	scoring rule list	catches high-pitched stress speech

2. Pattern label drift on monotone TTS

`06-monotone-dispatch` was scripted as monotone (Alex voice, low pitch base 45) but librosa returned pitch variance 36 — one point over the `> 35 ⇒ varied` threshold. Alex's natural prosody is more varied than the script intended. Real human monotone delivery sits well below this threshold; this is a TTS artifact.

3. Emotion classification is the weakest leg (50%)

Emotion depends on sentiment, which we synthesize from filename in this offline harness because librosa doesn't do sentiment. In live deployments with Deepgram's actual `sentiment_score`, expect noticeably better numbers — most fixtures with the right sentiment will land on the right emotion.

Notes & limitations

- Features were extracted with `librosa.pyin` rather than Deepgram. Deepgram's prosody.energy_mean is normalized differently, so live-deployed numbers will shift. This harness gives you a **regression baseline** for the rule logic itself.
- The 8 fixtures are TTS-generated (macOS `say`). Synthetic prosody is more uniform than human prosody, so accuracy looks better than on real human audio. For ground-truth accuracy, run a labeled corpus like RAVDESS or CREMA-D.
- The pass threshold of 50% is intentionally lax — this harness verifies the rules **fire as designed**, not that the design is optimal. Tune weights in `INFLECTION_RULES.md` if production data diverges.
