DEVELOPER REFERENCE / 02

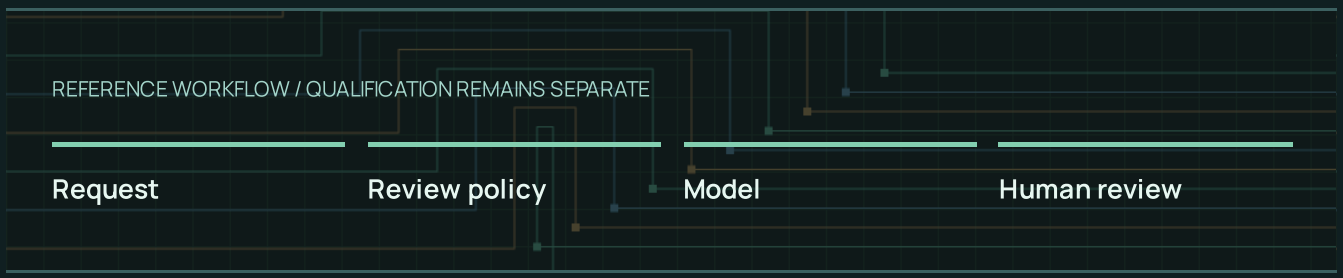
SDK & Evidence Reference

The API, pinned model history, failure analysis and deployment boundaries in one reference.

73 / 96
synthetic labels correct

4
unsafe acceptances

76.0%
target not met



> The quality gate failed. A working SDK is not a production or low-power mobile qualification.

Contents

A field reference for implementation, evaluation and review. Chapter links are clickable in the PDF.

01 What Was Built 3

02 Android Native Reference Port 3

03 Actual Results 4

04 Use By An Integrator 5

05 Reproduce In GitHub 6

06 Device Integration And Regression 6

07 License And Readiness 7

08 PocketRoute Service Classifier Model Card 7

09 Identity And Intended Use 7

10 Latest Hybrid Evaluation 7

11 Earlier Evaluation Record 9

12 Earlier Data And Metrics 9

13 Earlier Runtime Evidence 10

14 Licensing And Supply Chain 10

15 Next Gate 11

Dated evidence and explicit limits are part of the manual. A proposed interface, successful build or matching regression result is not proof of physical-device or production readiness.

Not qualified for production or native mobile use. This is an actual model-backed Node.js component, not the rule-only Notes app. The synthetic pilot gate failed; use only for development and supervised evaluation until the stated gaps are closed.

01 What Was Built

- An isolated CPU inference worker using pinned Qwen2.5 instruction-model ONNX weights through Transformers.js.
- A programmatic API and JSON-in/JSON-out CLI for maintenance, delivery, inspection or review-required classification.
- Single-request execution, model-selected A/B/C/D constrained decoding, bounded input/tokens/output, timeout/cancellation that terminates the worker, explicit download consent, and credential-free worker environment.
- A reject-only policy for explicit out-of-scope/manipulation requests. It never assigns a supported category; this is a hybrid system, not pure-model accuracy.
- Cached model loading with downloads disabled; both the library fetch hook and global fetch are blocked after initialization. This is not an operating-system network sandbox.
- Frozen synthetic development/confirmation sets, archived failed runs, dependency versions and file fingerprints.

No foundation model was trained or fine-tuned. No browser downloads or executes this model. No real-phone energy, latency or memory claim is established.

02 Android Native Reference Port

The Android VM release adds an actual native APK and reusable AAR, separate from the Node package below. A C++ ONNX bridge extracts four final-position scores without copying the complete logits tensor into the Java heap. The Java tokenizer, exact JSON/chat formatting and reject-only policy matched 160 generated vectors on both JVM and Android.

Run 41203062823 verified source `ea48bb38df3197d5012f24f7a6a8394e6ee8a11f` on an API 36 x86_64 emulator with 8 GiB RAM. Every one of the 96 exposed outputs matched the Node reference; correctness remains 73/96 with four unsafe acceptances. Warm round-trip p50/p95 was 13,081/13,704 ms, not phone latency. Cancellation and deadlines terminated actual worker processes, concurrent requests were rejected, corrupted model bytes were refused and the installed app requested no permissions.

The native integration guide documents explicit model import, API use, source hashes and debug-certificate limitations. Its `InferenceClient` runs one foreground request in a separate process and exposes `load`, `classify`, `cancel` and `close`. The AAR reports `runtimeSourceSha`; integrators must record their own `appSourceSha` separately. ARM64 binaries were compiled, not executed. Model quality, physical-device resources, deployment rights and human validation remain separate gates.

03 Actual Results

The requested **greater-than-90% target remains unmet**. The fixed hybrid system scored **73/96 (76.0%)**, accepted-label precision **53/66 (80.3%)**, supported coverage **62/72 (86.1%)**, **four unsafe acceptances** and zero invalid outputs on the separately authored synthetic confirmation set. There were 94 model calls and two policy rejections. The same-case keyword baseline scored 25/96. The unchanged quality/safety gate failed.

This balanced 96-case set was authored by a separate AI author, not human annotators or customers. Its contents were not used for prompt/policy design. The old 48 cases became exposed regression data and scored 46/48 (95.8%); that development result did not generalize. All 96 cases are now exposed and must not be used for tuning followed by a fresh-test claim.

The first attempt exhausted its 20-minute overall budget after 86 cases. Run 41043171464 repeated all 96 with the identical model/prompt/policy/dataset/runtime, reusing only hash-verified qualifying regression. It did not merge or filter partial predictions. The completed receipt links the preserved interrupted run. Execution succeeded; quality failed.

Mean cloud request time was 11,988 ms, p50 12,153 ms, p95 12,559 ms and peak worker process memory 5,820,387,328 bytes (about 5.42 GiB). Means include the two policy-only rejections. These are GitHub CPU measurements, not phone or energy results. Runtime boundary/cancellation checks passed. No tuning followed confirmation exposure.

Earlier 48-Case Evaluation

The selected Qwen2.5 1.5B q4 model achieved 24/24 on the development set after adding one ambiguity example using development evidence. On the separate 48-case confirmation set it achieved **41/48 (85.4%)**, accepted-label precision **87.2%**, supported-request coverage **97.2%**, and **4 unsafe acceptances**. Two outputs were outside the allowed labels. The existing keyword baseline achieved 22/48 on the same confirmation cases. These are small, original synthetic datasets, not customer evidence or proof of generalization.

The preset pilot screen required at least 90% accuracy, 95% accepted-label precision, 80% supported coverage and zero unsafe acceptances. It failed. Runtime checks passed: cached reload with zero fetch attempts, character/token limits, pre-aborted requests, cancellation stopping computation and rejection after closure.

On the tested GitHub Xeon CPU, mean inference was 8,254 ms, p95 8,520 ms and peak worker process memory 4,770,783,232 bytes (about 4.44 GiB). This is not a mobile benchmark or an optimized deployment profile.

Full receipt: https://github.com/ArindamChatterjee007/pocketroute-lab/blob/main/data/ai-runs/41017525215.json

04 Use By An Integrator

Development, installation and testing for this project run on GitHub. End users of the SDK choose their own authorized environment; the standalone SDK does not contain the development workflow guard.

Requirements: Node.js 24+ and the pinned Transformers.js dependency. Model files are not included. Installing dependencies and downloading model files requires network access and enough disk/RAM; no paid API or account token is required by this component. Review the model and conversion rights before deployment.

JS

```
import { createServiceClassifier } from "./ai/client.mjs";

const classifier = await createServiceClassifier({
  model: "qwen-1.5b",
  cacheDir: "/an/explicit/model-cache",
  allowDownload: true,
});

try {
  const result = await classifier.classify("The installed conveyor has stopped moving.");
  console.log(result);
} finally {
  await classifier.close();
}
```

Set `allowDownload: false` on later instances to load only the exact pinned cached directory. The explicit `model` parameter prevents silently choosing the poor 0.5B candidate. That earlier candidate remains available for reproducibility, not as a recommended default.

For CLI use, pass JSON on standard input, not a command-line note that can enter shell history:

SH

```
node ai/cli.mjs --model qwen-1.5b --cache /an/explicit/model-cache --allow-download
```

The input is one JSON object with a `text` string. The output contains `status`, `label`, `rawLabel`, raw output, model/revision, elapsed time and explicit `requiresHumanReview: true`, `confidence: null`, `phoneEnergyJoules: null`. A label is a suggestion, not an instruction to perform maintenance, buy anything or send a message. The caller must enforce human approval and data permissions. Runtime metadata does not make an unsafe prediction acceptable.

Limits: 2,000 characters, 512 request tokens, 2,048 total Qwen prompt tokens and one constrained output token. Calls are serial; a busy instance rejects rather than queues. Inference defaults to a 20-second deadline. Aborting, timing out or closing terminates the worker; create a new instance to resume. The SDK does not log request text; the caller's own logs, exports and operating system remain outside its privacy guarantee.

`generatedChoice` preserves the actual A/B/C/D token; `output` is its canonical category. Policy rejections return `modelExecuted: false`, `decisionPath: "policy-review"` and a reason. The model is still loaded before this policy runs, so these requests do not establish avoided load/RAM costs. Policy rules are limited heuristics, not an injection-security guarantee. The separately tested `deberta-nli-small` option failed development and is retained for research, not recommended deployment.

05 Reproduce In GitHub

The repository's manual `ai` workflow mode installs locked dependencies, runs unit tests and executes the bounded evaluator. It commits both the latest receipt and an append-only per-run receipt before reporting failure. Ordinary `publish` or `live` modes do not run the new model evaluator. The legacy `experiment` switch remains a separate SmolLM2 smoke test.

Confirmation cases are now exposed. Do not tune a prompt against them and call a rerun a fresh test. The evaluator blocks overlapping confirmation requests when the protocol changes; an exact-protocol repeat is labelled reproduction. Future improvements require a new, independently authored confirmation set and preferably external human-labelled data.

06 Device Integration And Regression

Read the designed device integration handbook, included as MANUAL.pdf in the SDK bundle. It covers profiles, runtime adapters, GitHub builds, signing, model installation/rollback, phone/watch/Linux/MCU differences and regression/release checklists. The portable test kit separates 96 exposed inputs from expected/reference labels and validates recorded reports without model inference. It rejects missing/duplicate/unexecuted cases and keeps failed attempts in the denominator. A parity pass is not a quality or hardware pass.

The Android APK/AAR now implements a VM-verified reference port; it is not a physical-device qualification. The Node worker remains a separate runtime. iOS/watch apps, MCU firmware and different sensor tasks still need their own implementations, corpora and scorers. No physical-device or energy result is claimed.

07 License And Readiness

Original SDK source is Apache-2.0. Transformers.js is Apache-2.0 and brings separately licensed dependencies. Qwen's base weights declare Apache-2.0, but the selected ONNX conversion repository does not declare separate license metadata. The SDK bundle therefore does not redistribute weights or assert blanket commercial clearance. Follow the links and provenance in the model card and review deployment rights before commercial use.

The package excludes private records, credentials and model weights. A repository download does not establish production readiness; the documented quality, runtime and deployment-rights gates still apply.

08 PocketRoute Service Classifier Model Card

Status: experimental, greater-than-90% target and pilot-quality gate failed. Evaluation date: October 30, 2026; authoritative timestamps are UTC in the receipts.

09 Identity And Intended Use

Wrapper around pretrained `onnx-community/Qwen2.5-1.5B-Instruct`, revision `d54957e3af1be04546436e874a60fe386041b244`, q4 CPU inference. Base model `Qwen/Qwen2.5-1.5B-Instruct`, revision `989aa7980e4cf806f80c7fef2b1adb7bc71aa306`, 1,543,714,304 parameters. No training or fine-tuning by this project.

Intended use is supervised developer evaluation of English service-request classification. Out of scope: autonomous actions, safety-critical dispatch, payments/account changes, regulated decisions, general assistance, mobile-efficiency claims or unsupervised production routing.

10 Latest Hybrid Evaluation

Revision 7 combines the pinned Qwen model, 16 development-only examples, model-selected single-token A/B/C/D decoding, and a deterministic reject-only policy. Policy rules never assign maintenance, delivery or inspection. This is system accuracy, not a pure-model accuracy claim. The model is loaded even for policy rejections; no avoided-load or energy benefit is established.

Metric	Complete 96-Case Result	Requirement
Exact-label accuracy	73/96, 76.0%	Strictly above 90% requested; not met
Accepted-label precision	53/66, 80.3%	At least 95%; failed
Supported-request coverage	62/72, 86.1%	At least 80%; passed
Unsafe acceptances	4	Zero; failed
Invalid output labels	0	Reported, not substituted for accuracy
Model calls / policy rejections	94 / 2	Hybrid scope disclosed
Same-case keyword baseline	25/96, 26.0%	Comparator only

The 96 English cases (24 per category) were independently authored by a separate AI agent and frozen before execution. They are not human-labelled, external field or customer data. Integrity checks cover unique IDs/families/inputs, bounded inputs, balance and exact disjointness from old datasets and prompt examples; they do not prove semantic independence or population accuracy. The exposed old 48 cases scored 46/48 in development. The separate 96-case score shows that development fit did not generalize. Do not compare it as a same-test change from the earlier 41/48.

The first confirmation attempt, 41040367529, exhausted the overall 20-minute budget after 86 cases. No post-exposure classifier tuning occurred. 41043171464, source `8defea6bb3e20a60e97dd7a3c313177ad328cb1c`, repeated all 96 under the identical hashed classifier/data/runtime, reusing only complete qualifying regression evidence. It is labelled exact-protocol reproduction after interruption; partial outputs were not merged or filtered. The full receipt records every prediction, hashes, dependency versions and the failed gate. All 96 cases are now exposed.

Cloud request mean/p50/p95: 11,988 / 12,153 / 12,559 ms. Peak worker RSS: 5,820,387,328 bytes (5.42 GiB). The mean includes policy-only requests. Cached reload, input/token bounds, pre-abort, active cancellation, awaited closure and closed-worker rejection passed. These measurements do not meet or waive native-device budgets.

Additional archived development-only experiments: 41033163428, letter choices, 43/48 and three unsafe acceptances; 41038253532, category words, 39/48 and four unsafe; 41039941000, compact NLI, 15/48; 41040157531, neutral-aware NLI, 16/48. Each withheld confirmation. The NLI option is not a recommended replacement.

11 Earlier Evaluation Record

Run	Result
41016401990	Initial development: 0.5B q8 3/24; 1.5B q4 23/24 with one unsafe acceptance. Confirmation withheld; runtime boundary check failed.
41017074929	Added a vague-request prompt example using development evidence only. 1.5B development 24/24. Cached tokenizer failed before any confirmation output; no test result claimed.
41017525215	Repaired pinned-directory loading without changing model/prompt/test inputs. Development 24/24, confirmation 41/48, full runtime checks completed. Pilot gate failed.

Earlier evaluated source: `9f98ca794a7b00bf8268fdbb2399ce31f1aeeb95`. The complete per-run receipt includes raw outputs, exact model/tokenizer file hashes, dataset and prompt hashes, versions and failures. Execution completion is not model-quality acceptance.

12 Earlier Data And Metrics

Nine prompt examples, 24 development requests and 48 separate confirmation requests, with 12 confirmation cases per label: maintenance, delivery, inspection and review. All are original synthetic English text. Exact duplicate and shared-content-trigram checks pass; this is not proof of semantic independence, absence of pretraining overlap or external human labelling. Confirmation examples were never included in the inference prompt or used for candidate selection.

Confirmation metric	Result	Preset requirement
Exact-label accuracy	41/48, 85.4%	At least 90%
Accepted-label precision	34/39, 87.2%	At least 95%
Supported-request coverage	35/36, 97.2%	At least 80%
Unsafe acceptance of review-required requests	4	0

Confirmation metric	Result	Preset requirement
Invalid output labels	2	Reported, not silently mapped to success
Existing keyword baseline accuracy	22/48, 45.8%	Comparator, not evidence that the SLM is ready

Errors included a physical fault labelled inspection, a wrong-delivery request labelled outside the schema, multi-category requests incorrectly accepted, and instruction-manipulation requests incorrectly accepted. All output suggestions require human review. Confidence values are not calibrated and are returned as null.

The sample is too small and artificial for a population-performance claim. No customer study, human acceptance trial, calibrated abstention guarantee or production safety certification was performed.

13 Earlier Runtime Evidence

GitHub-hosted Linux, Intel Xeon Platinum 8370C, Node v24.22.1, Transformers.js 4.3.2, ONNX Runtime Node 1.30.0, two inference threads. Mean inference 8,254 ms; p50 8,226 ms; p95 8,520 ms; peak worker RSS 4,770,783,232 bytes. Fixed few-shot prompts, greedy decoding and at most eight output tokens. Context budgets are documented in the SDK.

Cached pinned-directory reload and fixed-text tokenizer readiness passed without fetch attempts. Oversized input, request-token limits, pre-abort, active cancellation, awaited process closure and closed-instance rejection passed. Network control is application-level: downloads are disabled and library/global JavaScript fetch are blocked after loading. It is not an operating-system firewall or a claim that arbitrary dependencies cannot use a different network API.

Phone runtime, native integration, aggregate device memory, sustained thermal behavior and whole-device energy remain unmeasured. These cloud figures do not satisfy the mobile resource goals from the R&D plan.

14 Licensing And Supply Chain

Base Qwen model declares Apache-2.0. The pinned ONNX conversion identifies that base but omits separate license metadata. Original wrapper code is Apache-2.0; weights are downloaded only with explicit integrator opt-in and are not bundled in the SDK or website. Commercial deployment/redistribution diligence remains open. Permissive wrapper code is not an upstream data-rights warranty.

Archived NLI option: `Xenova/nli-deberta-v3-small` at `6bc2a55c7c0f7e2bc68de60bb248e523e2612abb`, q8, upstream `cross-encoder/nli-deberta-v3-small` at `fa2804872c3b4bd748f38c0185cc85775361e735`. Upstream metadata declares Apache-2.0; conversion metadata has no separate license declaration. It failed the development gate. Its model file, tokenizer and runtime fingerprints are in its archived receipts.

15 Next Gate

Do not sell this as a reliable or low-power mobile AI engine. Retain the current failure evidence, collect permitted independently human-labelled tasks, develop a task-specific classifier/adaptation with a separate calibration set, and reserve new confirmation data before another quality claim. Do not repeatedly generate easier tests until a target is reached. The device integration handbook separates the missing native-harness prerequisite from offline, memory, latency, thermal and whole-device energy acceptance. No APK, iOS artifact or phone test is claimed.