

flutter_kiwi_nlp: A Native-First, Cross-Platform Korean NLP Plugin for Flutter

Jai-Chang Park*

February 18, 2026

Abstract

This paper presents `flutter_kiwi_nlp`, a production-oriented Flutter plugin for Korean morphological analysis built on Kiwi. The package exposes one stable Dart API while internally operating two runtime stacks: Native (Dart FFI + C bridge + Kiwi shared library) and Web (Dart JS interop + kiwi-nlp WASM). This design enables a single application codebase across Android, iOS, macOS, Linux, Windows, and Web. Unlike ONNX-export deployment paths, this integration does not require adding an extra generic inference runtime layer because Kiwi already provides the analysis engine and model format used here.

The implementation is explicitly aligned with on-device AI requirements: local inference execution, reduced text egress by default, predictable latency without per-request network dependence, and operational fallback controls for model provisioning. Empirically, repeated benchmark trials show a *warm-path throughput reversal* (steady-state path, initialization excluded): Flutter warm throughput can exceed the Python-native baseline: 3.12x on desktop, 3.87x on iOS simulator (debug), and 1.52x on Android emulator (release). Added Linux/Windows CI rows show clear execution-mode sensitivity: single mode is 0.64x of Python throughput on both OSes, while batch mode is 1.54–1.55x faster than Python on the same runners. The updated benchmark pipeline also adds boundary-decomposed measurements (pure processing vs full JSON path), quantifying serialization/parsing overhead at 13.89–47.93% depending on runtime profile. This warm-path reversal does not imply end-to-end session superiority: cold-start initialization and short-session effective-throughput remain weaker than the Python baseline on all measured platforms. Cross-runtime linguistic agreement remains close on gold corpora (88.39% vs 88.58% token agreement; 84.90% vs 85.55% POS agreement). On native targets, inference can run fully offline after one-time model provisioning.

This paper provides an implementation-complete, reproducible specification of API contract, runtime parity rules, build automation hooks, security boundaries, failure taxonomy, benchmark protocol, and quantified ecosystem survey signals.

1 Introduction

Korean morphological analysis is a foundational primitive for search, retrieval, classification, and generation workflows. In Flutter environments, application logic can be shared across targets, but language runtime integration remains platform-specific. `flutter_kiwi_nlp` addresses this mismatch by encapsulating runtime diversity behind a single API boundary.

The central engineering problem is not only to expose Kiwi functionality, but to make runtime behavior predictable when deployment environments differ in:

*Google Developer Expert (GDE), Dart & Flutter; GDG (Google Developer Groups) Golang Korea; Flutter Seoul.

- native dynamic library semantics,
- browser module and WASM loading constraints,
- model-file availability,
- platform build toolchains and artifact formats.

1.1 On-Device AI Positioning

This plugin is designed as an on-device AI integration layer, not a network-first inference client. In this context, "on-device" means that tokenization and morphological analysis execute inside the host application process (native FFI path) or in-browser runtime (WASM path), with optional model download only for provisioning. This positioning provides:

- stronger default data locality for analyzed text,
- lower dependency on network availability during inference,
- deterministic runtime behavior under explicit model/version controls,
- easier integration into privacy- or compliance-sensitive workloads.

For native targets, once model artifacts are bundled or cached locally, analysis can execute without network connectivity (offline-first inference path). For web targets, comparable offline behavior requires deployment-specific caching and self-hosting policy because the default bootstrap uses CDN module/WASM URLs.

1.2 Contribution Type

This paper is positioned as a **systems/resource** contribution rather than an algorithmic NLP paper. Its primary value is integration engineering: cross-runtime API unification, deterministic fallback policy, reproducible benchmark protocol, and deployable build/runtime controls for Flutter applications. This revision also contributes glass-box performance profiling via boundary-decomposed metrics (pure vs full path) and cross-OS CI benchmark rows that expose execution-mode sensitivity. It does not claim a new morphological decoding algorithm or a new Korean language model architecture.

2 Related Work

This revision expands the related-work coverage to include core Transformer foundations, compact/efficient variants, on-device inference studies, and Korean-language-specific resources. For foundation and compression lineage, prior work includes BERT, ALBERT, ELECTRA, DistilBERT, TinyBERT, MiniLM, and MobileBERT [1, 2, 3, 4, 5, 6, 7]. These studies establish the main trade-off surface between representational quality and model size/latency.

For edge execution and deployment-oriented efficiency, SqueezeBERT, EdgeBERT, and DeeBERT provide architectural or runtime pathways for lower-latency transformer inference, while pNLP-Mixer explores an all-MLP design for compact on-device language modeling [8, 9, 10, 11]. Application focused on-device work reports practical workloads such as VQA, form filling, and smart-reply code-switching, and personalization-oriented studies analyze device-local vocabulary adaptation under memory/latency constraints [12, 13].

For Korean NLP context, KR-BERT, KLUE, KoBigBird-large, and character-level Korean morphological analysis/POS tagging provide complementary evidence on language-specific tokenization, evaluation, and modeling behavior [14, 15, 16, 17]. These works are primarily model or benchmark contributions; by contrast, this paper focuses on systems integration for production Flutter apps, including runtime selection, artifact provisioning, fallback semantics, and reproducibility constraints.

The Korean morphological analyzer ecosystem itself is also important related work context. Widely used dictionary-centric analyzers and lexicons such as MeCab-ko and MeCab-ko-dic represent a practical baseline in many production pipelines, while Khaiii represents a neural approach with different quality and runtime trade-offs [18, 19, 20]. In Python workflows, KoNLPy is often used as an integration layer over multiple Korean analyzers, affecting how researchers compare or operationalize analyzers in practice [21]. Relative to that landscape, this paper’s main contribution is not proposing a new Korean analyzer itself, but delivering a native-first, cross-platform Flutter integration path for Kiwi with explicit reproducibility and deployment controls.

Table 1 summarizes commonly used Korean morphological analyzers/toolkits in practice, with repository-level adoption signals collected on February 17, 2026 via GitHub API snapshots [22].

Table 1: Representative Korean morphological analyzers and usage landscape (snapshot: 2026-02-17)

Analyzer / Toolkit	Primary runtime	Typical usage context	Adoption signal (snapshot)	Source
Kiwi	Native C++ (+ wrappers)	Production morphology/POS pipelines; offline embedding	GitHub stars: 671	[23]
MeCab-ko (+ mecab-ko-dic)	Native C++	Established tokenization/POS baseline; commonly used via KoNLPy Mecab path	MeCab upstream + Eunjeon lineage references; no canonical active GitHub org repo in this snapshot	[18, 19, 21]
Khaiii	Native C++/Python wrapper	Neural Korean analyzer for batch/server usage	GitHub stars: 1,448	[20]
Open Korean Text (Okt)	JVM/Scala (+ wrappers)	Social-text normalization/tokenization in JVM and Python workflows	GitHub stars: 656	[24, 21]
KOMORAN	JVM	Java production pipelines and KoNLPy-integrated experiments	GitHub stars: 311	[25, 21]
KoNLPy (toolkit)	Python wrapper hub	Unified interface over Kkma, Hannanum, Komoran, Mecab, Okt	GitHub stars: 1,486	[21, 26]

For benchmark methodology, GLUE and SuperGLUE establish widely reused task-oriented NLU evaluation structure, and Long Range Arena targets efficiency comparison under long-context settings [27, 28, 29]. For systems-level performance reporting, DAWNBench and MLPerf Inference provide complementary protocol ideas such as time-to-accuracy framing and cross-stack inference benchmarking [30, 31].

From a systems perspective, on-device ML discussions increasingly emphasize data locality, reduced dependency on always-on connectivity, and deployment practicality under edge constraints [32]. These themes align with the operational goals of `flutter_kiwi_nlp`: local inference path, deterministic runtime controls, and explicit provisioning/fallback mechanisms.

Privacy-preserving distributed training literature such as FedAvg also informs the broader motivation for keeping user data local when possible [33]. Although this plugin targets inference rather than training, the same locality principle reinforces its offline-first native execution model.

3 Background: Flutter and Kiwi

3.1 What Flutter Is

Flutter is an open-source UI toolkit for building applications from one Dart codebase across mobile, desktop, and web targets. Its rendering and widget model allows large parts of application logic to be shared, but low-level platform integration is still target-specific.

For systems like NLP analyzers, Flutter integration typically requires one of:

- platform channels (message-based bridge), or
- FFI plugins (direct native library interop from Dart).

`flutter_kiwi_nlp` uses FFI for native platforms and JS interop on web, which is why the plugin architecture is more complex than standard UI-only Flutter packages.

3.2 What Dart Is and Why It Was Created

Dart is an open-source, client-optimized language developed by Google and used as the language foundation of Flutter. The Dart project positions the language goal as productive multi-platform development paired with a flexible runtime platform.¹

The practical reason this matters for plugin engineering is that Dart is designed for both development-time velocity and production deployment across multiple backends. In current official language positioning, this includes:

- fast iterative development workflows (for example, hot-reload-oriented tooling in Flutter),
- ahead-of-time native compilation for device/desktop targets, and
- web-target compilation paths (JavaScript and WebAssembly).²

For this plugin, these Dart properties directly motivate the design choice to publish one stable Dart API while internally dispatching to platform-specific runtime backends.

3.3 What Rust Is and Why It Is Mentioned

Rust is an open-source systems programming language designed around memory safety, strong compile-time guarantees, and predictable performance without a garbage collector.³

This repository is not primarily implemented in Rust: native runtime integration here is built on a Dart FFI layer and a C bridge that loads Kiwi artifacts. Rust is still worth documenting in this paper because reviewers often ask about Rust’s role in modern WebAssembly ecosystems. In this plugin, consumer builds do not require a Rust toolchain; web execution depends on the distributed `kiwi-nlp` JS/WASM artifacts and native execution depends on platform libraries. This is an intentional interoperability tradeoff: the package prioritizes adopting upstream Kiwi distribution artifacts over introducing an additional language toolchain requirement for plugin consumers.

¹<https://dart.dev/overview>

²<https://dart.dev/>

³<https://www.rust-lang.org/>

3.4 What WebAssembly (WASM) Is

WebAssembly (WASM) is a compact binary instruction format and execution model for stack-based virtual machines.⁴ On the web it runs inside the browser sandbox, typically loaded by JavaScript bootstrap code, and enables near-native computational kernels while preserving browser security constraints.

In plugin context, WASM is not a replacement for Flutter itself; it is a backend execution target used by the web runtime path to execute Kiwi analysis logic in browser environments.

3.5 What Kiwi Is

Kiwi is a Korean morphological analysis engine distributed primarily as a native library and ecosystem artifacts. In this repository, Kiwi is consumed through:

- native dynamic libraries loaded by a C bridge on Android/iOS/macOS/Linux/Windows,
- the `kiwi-nlp` JavaScript/WASM package on web.

Operationally, Kiwi performs segmentation and POS-tagged token analysis, and the plugin exposes that capability as: `create`, `analyze`, `addUserWord`, and `close`.

Model files are external artifacts (for example `cong.mdl`, `default.dict`, `typo.dict`) that must be present locally, packaged as Flutter assets, or obtained through archive fallback.

3.5.1 How Kiwi Models Are Trained (Upstream Summary)

Kiwi’s upstream design separates dictionary/rule-driven candidate generation from statistical disambiguation. In practical terms, this means model training is centered on language-model estimation over morphologically analyzed corpora, while lexicon/rule components remain explicit resources [23, 34].

Upstream documentation and the Kiwi paper describe training data composition using large Korean morphological corpora, including Sejong and National Institute of Korean Language resources, and report model-family evolution from KNLM-focused scoring to Skip-Bigram and contextual n-gram embedding variants [23, 34].

For `flutter_kiwi_nlp`, the operational implication is explicit: this package consumes pre-built Kiwi model artifacts for inference, and does not re-implement Kiwi training in the Flutter build/runtime path.

3.5.2 Kiwi Internal Architecture (From Upstream Source)

Upstream Kiwi implementation and public headers indicate that runtime decoding is organized as a morphology candidate graph/lattice search with language-model scoring, rather than a Transformer encoder stack [23, 35, 36, 37]. At the API/type level, model families are explicitly defined as: `knlm` (Kneser–Ney LM), `sbg` (Skip-Bigram), and `cong/congGlobal` (contextual N-gram embedding LM variants), with no Transformer model type in the exposed enum [35].

At inference-time, Kiwi keeps LM state while evaluating candidate paths over the morpheme graph and applies additional rule-based penalties/bonuses for morphological compatibility and punctuation/quote state handling [36, 37]. This is closer to a graph-decoding + statistical language-model architecture than to deep self-attention sequence encoding.

⁴<https://webassembly.org/>

3.5.3 Conceptual Decoding Objective

For reviewer readability, the upstream behavior can be summarized as a lattice path optimization problem (conceptual/illustrative form; not a claim of reproducing every internal constant/feature exactly):

$$\pi^* = \arg \max_{\pi \in \Pi(G)} S(\pi),$$

where G is the morpheme lattice and $\Pi(G)$ is the set of valid paths. The path score is composed additively:

$$S(\pi) = \sum_{t=1}^{T(\pi)} \left[s_{\text{lm}}(e_t, h_{t-1}) + \lambda^\top \phi_{\text{rule}}(e_t, c_t) + \gamma^\top \phi_{\text{lex}}(e_t) \right].$$

Here, e_t denotes the selected candidate edge at step t , h_{t-1} is the LM state carried from prior steps, c_t encodes local context such as quote/punctuation conditions, and $\phi_{\text{rule}}, \phi_{\text{lex}}$ represent rule/lexicon-derived feature contributions.

State progression and pruning can be abstracted as:

$$h_t = F(h_{t-1}, e_t), \quad \mathcal{B}_t = \text{TopK}(\mathcal{P}_t(G), K_{\text{beam}}, S),$$

where $\mathcal{P}_t(G)$ is the partial-path set at depth t , and \mathcal{B}_t is the retained beam-like frontier with width K_{beam} .

The LM-family switch exposed by Kiwi model type can be written as:

$$s_{\text{lm}} = \begin{cases} s_{\text{knlm}}, & m = \text{knlm}, \\ s_{\text{knlm}} + \Delta_{\text{sbg}}, & m = \text{sbg}, \\ s_{\text{cong}}, & m = \text{cong}, \\ s_{\text{cong-global}}, & m = \text{congGlobal}. \end{cases}$$

This formulation explains why Kiwi behaves as graph decoding with explicit LM state and rule scoring, rather than Transformer self-attention inference.

3.5.4 Skip-Bigram and Contextual N-gram Scoring (Conceptual)

Because the plugin references Kiwi model families `sbg` and `cong/congGlobal`, it is useful to provide explicit conceptual scoring forms for reviewer intuition:

$$\Delta_{\text{sbg}}(e_t, h_{t-1}) \approx \sum_{k=1}^{\min(K_{\text{skip}}, t-1)} \alpha_k \psi_k(m_t, m_{t-k}),$$

where m_t is the morpheme realized by edge e_t , k is skip distance, and α_k is a distance-dependent weight. The function ψ_k denotes skip-bigram interaction score components. Here K_{skip} is the maximum skip order considered.

A generic contextual n-gram embedding style score can be expressed as:

$$c_t = \sum_{j=1}^{\min(n-1, t-1)} W_j v(m_{t-j}), \quad s_{\text{cong}}(e_t, h_{t-1}) \approx u(m_t)^\top c_t + b(m_t),$$

where $v(\cdot)$ and $u(\cdot)$ denote context/target embeddings and W_j encodes positional/context projection.

For the long-context variant, a practical abstraction is:

$$c_t^{\text{global}} = \sum_{j=1}^{\min(n_g-1, t-1)} W_j^{(g)} v(m_{t-j}), \quad s_{\text{cong-global}}(e_t, h_{t-1}) \approx u(m_t)^\top c_t^{\text{global}} + b_g(m_t),$$

with $n_g > n$ to reflect wider context coverage.

These equations are intentionally presented as conceptual abstractions to explain model-family behavior at paper level; exact implementation constants and feature composition are defined by upstream Kiwi internals and model artifacts. They are explanatory model-family formulations and are not used directly to compute the benchmark tables in Section 17.

Figure 1 summarizes this upstream-oriented view of Kiwi internals as used in the plugin context.

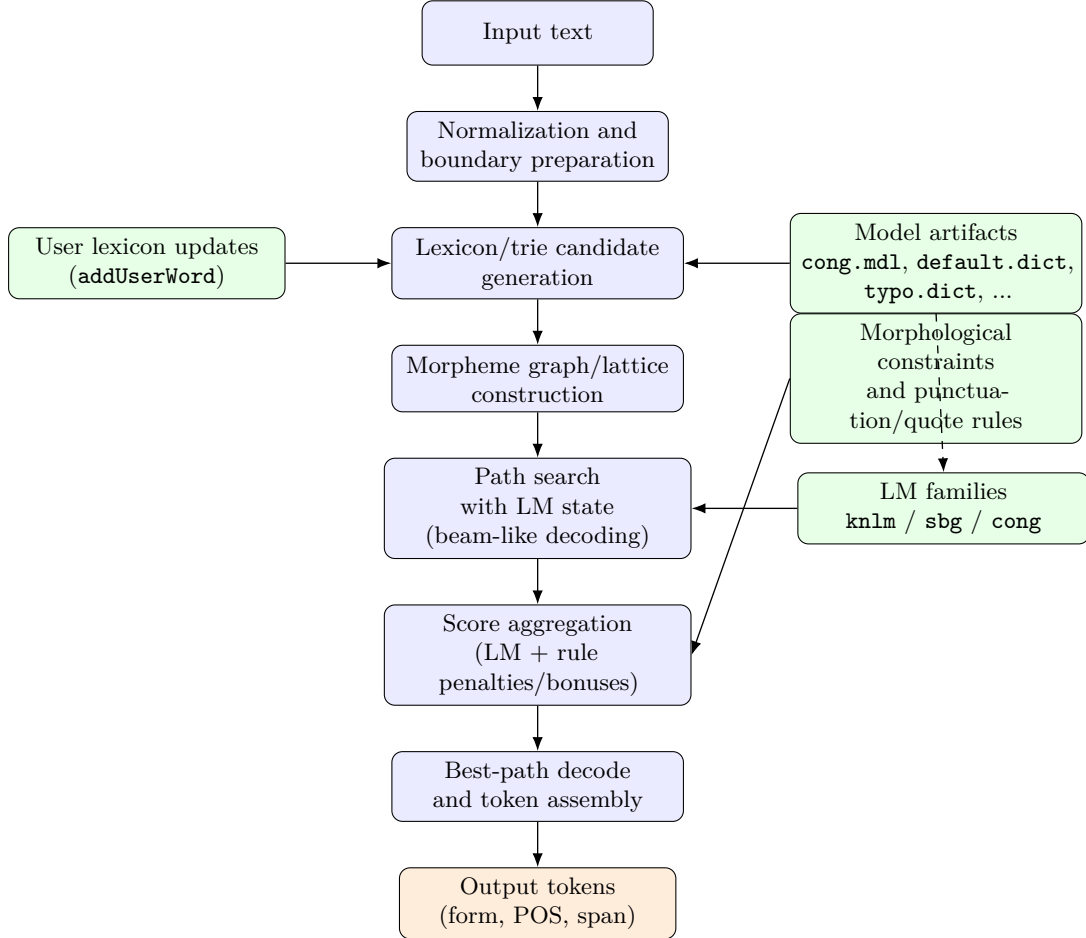


Figure 1: Conceptual Kiwi internal architecture based on upstream source inspection.

3.5.5 Why This Is Not a Transformer

Transformer-based analyzers typically center on stacked self-attention blocks and dense neural sequence representations. Kiwi’s upstream architecture, by contrast, is designed around:

- lexicon/trie-driven candidate generation and morphological constraints,
- explicit path search with LM-state progression over candidates,

- lightweight N-gram-family scoring models (`knlm/sbg/cong`).

This distinction explains why Kiwi integration characteristics differ from Transformer runtimes: lower footprint and predictable CPU inference behavior are prioritized, while broad semantic representation power is not the primary design target in this component.

3.5.6 Why No ONNX Runtime Layer Is Required

Some modern NLP deployment stacks package neural models in ONNX format and ship an additional generic runtime engine (for example, ONNX Runtime) inside the application process. `flutter_kiwi_nlp` intentionally does not add that extra layer, because Kiwi already provides its own inference/decode engine and model format for Korean morphological analysis.

This design choice has practical implications:

- fewer runtime components to integrate and version-lock,
- simpler build/runtime dependency surface for plugin consumers,
- no separate ONNX operator/runtime compatibility matrix to maintain.

Important scope clarification: this is a dependency and integration argument, not a universal speed claim against all ONNX-based analyzers. In this paper’s own benchmark, `flutter_kiwi_nlp` can outperform Python-native `kiwipiepy` on warm-path throughput under the latest measured profile, but still shows slower cold start and lower short-session effective throughput because initialization and bridge costs are front-loaded.

3.6 Why This Integration Is Non-Trivial

The plugin is not a direct wrapper around one runtime. It is an orchestration layer that has to keep semantics stable across different execution models:

- native C ABI with explicit memory ownership,
- browser JS/WASM with promise-based async semantics,
- platform-specific build systems and artifact formats.

4 Prior Ecosystem Survey: Existing Kiwi Wrappers

4.1 Survey Scope and Method

This paper includes an explicit prior-ecosystem survey to position `flutter_kiwi_nlp` against existing Kiwi bindings. The survey source is the upstream Kiwi repository README and linked binding projects, plus GitHub repository metadata snapshots collected on February 17, 2026.⁵

4.2 Observed Upstream Wrapper/Binding Landscape

Table 2 summarizes the wrappers and binding channels explicitly referenced upstream.

⁵<https://github.com/bab2min/Kiwi>

Table 2: Upstream Kiwi wrapper/binding survey (as documented by Kiwi)

Category	Location	Notes
C API	<code>include/kiwi/capi.h</code>	Core C interface for native embedding.
Prebuilt binaries	Kiwi releases page	Windows/Linux/macOS/Android library and model artifacts are distributed through release assets.
C# wrapper (official GUI usage)	https://github.com/ab2min/kiwi-gui	Upstream points to C# wrapper used by official GUI tooling.
C# wrapper (community)	https://github.com/X3exp/NetKiwi	Community-contributed multiplatform C# wrapper linked by upstream README.
Python wrapper	https://github.com/ab2min/kiwipiepy	Officially documented Python3 API package.
Java binding	<code>bindings/java</code> in Kiwi repository	Java 1.8+ binding path documented upstream.
Android library	Kiwi release asset (<code>kiwi-android-VERSION.aar</code>)	Android NDK-based AAR distribution path documented upstream.
R wrapper	https://mrchypark.github.io/elbird/	Community-contributed R wrapper linked by upstream README.
Go wrapper	https://github.com/odingpot/kiwigo	Community Go wrapper linked by upstream README.
WebAssembly binding	<code>bindings/wasm</code> in Kiwi repository	JavaScript/TypeScript-facing WASM binding path documented upstream.

4.3 Gap Analysis and Motivation for This Work

The upstream survey shows broad language coverage around Kiwi, but it also shows a practical integration gap for Flutter package consumers:

- no upstream-listed first-class Dart/Flutter wrapper entry,
- no upstream-listed packaging path that unifies native (mobile/desktop) and web (WASM) behind one Dart API contract,
- no upstream-listed Flutter-specific build-hook automation for shipping platform artifacts in plugin workflows.

This gap is the direct motivation for `flutter_kiwi_nlp`: one Flutter-native package that preserves Kiwi backend capability while adding runtime abstraction, model resolution policy, and target-specific packaging automation needed by real Flutter deployments.

4.4 Quantified Maintenance Signals

To reduce purely descriptive bias, this revision adds repository-level maintenance signals collected from GitHub REST API snapshots on February 17, 2026 (`tool/benchmark/collect_wrapper_activity.py`). Table 3 reports latest release date, latest commit date, and commit velocity windows for each wrapper repository.

Table 3: Quantified wrapper maintenance signals (as-of 2026-02-17)

Wrapper	Repo	Latest release	Last commit	90d commits	365d commits	Stars
kiwi-gui	bab2min/kiwi-gui	2025-12-22 (57d)	2025-12-22 (57d)	10	18	14
NetKiwi	EX3exp/NetKiwi	N/A	2025-02-18 (364d)	0	1	2
kiwipiepy	bab2min/kiwipiepy	2025-12-15 (64d)	2025-12-25 (54d)	33	114	357
Kiwi Java binding	bab2min/Kiwi	2025-12-15 (64d)	2026-02-02 (15d)	30	279	671
Kiwi WASM binding	bab2min/Kiwi	2025-12-15 (64d)	2026-02-02 (15d)	30	279	671
elbird	mrchypark/elbird	2025-12-30 (49d)	2025-12-30 (49d)	37	44	34
kiwigo	codingpot/kiwigo	N/A	2025-10-15 (125d)	0	5	30

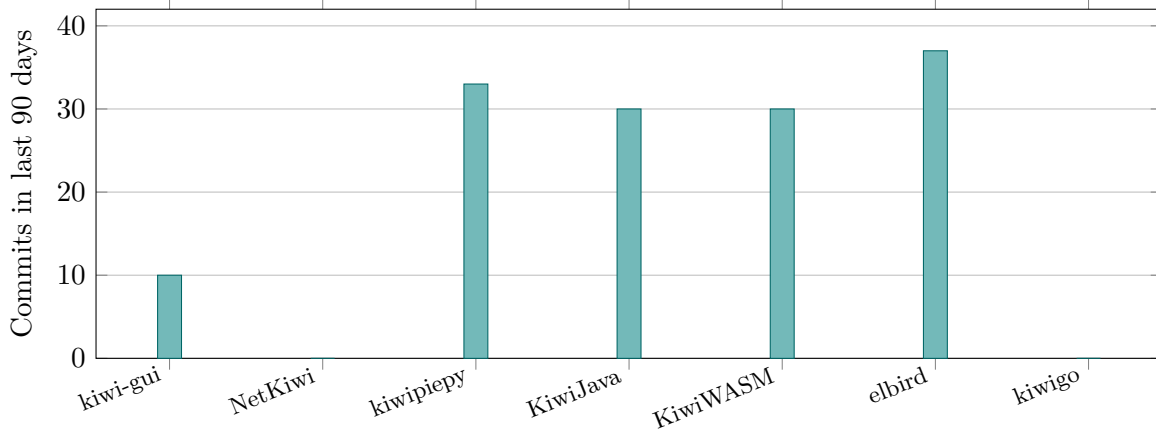


Figure 2: Recent wrapper activity (90-day commit count).

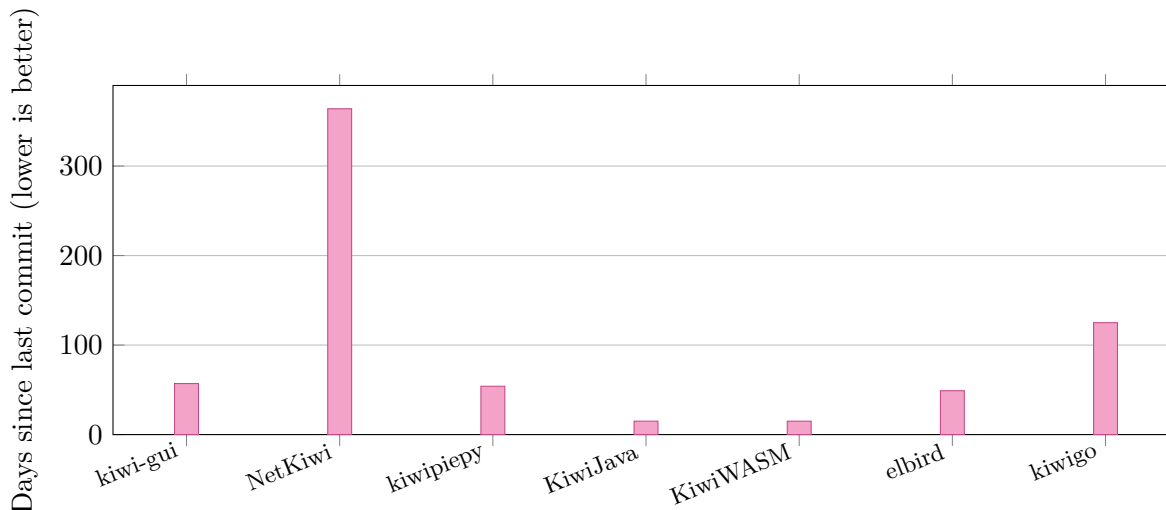


Figure 3: Repository recency profile for upstream wrappers.

4.5 Survey Evidence Limits

These maintenance signals are repository-level proxies. For Java and WASM, the metrics come from the shared Kiwi monorepo and do not isolate directory-level binding effort. Issue response latency and download velocity are also not yet included.

5 Design Goals and Non-Goals

5.1 Primary Goals

1. **Stable API contract:** identical Dart signatures across native, web, and unsupported stubs.
2. **Operational resilience:** layered model resolution, archive fallback, and explicit error propagation.
3. **Build-time ergonomics:** automatic preparation of missing Kiwi artifacts during platform builds.
4. **Typed outputs:** deterministic output schema (result \rightarrow candidates \rightarrow tokens).

5.2 Explicit Non-Goals

- Novel morphological-analysis algorithm or model-training contribution.
- Full feature parity with every upstream Kiwi API detail.
- Guaranteed throughput parity with Python-native `kiwipiepy`.
- Support for targets not declared in Flutter plugin metadata (for example, Fuchsia).

6 Supported Platform Matrix

Table 4 summarizes currently supported targets, execution backend, artifact strategy, and packaging hook.

Table 4: Current platform support matrix

Platform	Flutter plugin declaration	Runtime backend	Bundled artifact	Preparation hook
Android	Supported (<code>ffiPlugin</code>)	Native FFI + C bridge + <code>libkiwi.so</code>	<code>android/src/main/jniLibs/abi- arm64-v8a/libkiwi.so</code>	Gradle task <code>prepareKiwiAndroidLibs</code> bound to <code>preBuild</code>
iOS	Supported (<code>ffiPlugin</code>)	Native FFI + C bridge + Kiwi framework	<code>ios/Frameworks/Kiwi.xcframework</code>	Podspec hook (<code>prepare_command</code>) runs <code>tool/build_ios_kiwi_xcframework.sh</code>
macOS	Supported (<code>ffiPlugin</code>)	Native FFI + C bridge + <code>libkiwi</code>	<code>macos/Frameworks/libkiwi.dylib</code>	Podspec hook (<code>prepare_command</code>) runs <code>tool/build_macos_kiwi_dylib.sh</code>
Linux	Supported (<code>ffiPlugin</code>)	Native FFI + C bridge + <code>libkiwi.so</code>	<code>linux/prebuilt/libkiwi.so</code>	CMake custom target <code>prepare_kiwi_linux_lib</code>
Windows	Supported (<code>ffiPlugin</code>)	Native FFI + C bridge + <code>kiwi.dll</code>	<code>windows/prebuilt/kiwi.dll</code>	CMake custom target <code>prepare_kiwi_windows_dll</code>
Web	Supported (web plugin class)	JS interop + <code>kiwi-nlp</code> WASM	Model files loaded by URL or archive bytes in memory	Runtime module/model loader in <code>kiwi_analyzer_web.dart</code>
Fuchsia	Not declared	Stub behavior	N/A	N/A

6.1 Architecture Coverage

- Android ABIs default to `arm64-v8a` and `x86_64`.
- iOS framework build includes `iphoneos arm64` and simulator `arm64/x86_64` slices.
- macOS default build targets `arm64` and `x86_64`.
- Linux supports host-driven architecture mapping (`x86_64`, `arm64`, `ppc64le`) in build scripts.
- Windows arch normalization supports `x64 (x86_64)`, `Win32 (x86)`, and `arm64`; prebuilt download path currently covers `x64/Win32`.

6.2 Artifact Footprint Snapshot

Table 5 summarizes build/model size signals that are already tracked in project documentation and re-checked from workspace artifacts for this revision. These are engineering reference numbers, not

final store-delivery APK/IPA install sizes.

Table 5: Artifact footprint snapshot (workspace measurement, 2026-02-18)

Artifact	Size (approx.)	Notes
Default model directory (<code>assets/kiwi-models/cong/base</code>)	99,308,057 bytes (≈ 94.71 MiB)	Uncompressed model files used by the plugin.
Same default model compressed as local .tgz archive	79,494,329 bytes (≈ 75.8 MiB)	Local compression reference from this workspace.
<code>android/src/main/jniLibs/arm64-v8a/libkiwi.so</code>	166,229,088 bytes (≈ 158.53 MiB)	Current artifact is with <code>debug_info</code> , not stripped .
<code>android/src/main/jniLibs/x86_64/libkiwi.so</code>	200,071,656 bytes (≈ 190.80 MiB)	Current artifact is with <code>debug_info</code> , not stripped .

To avoid ambiguity between source artifacts and packaged binaries, Table 6 reports example Android app outputs for both debug and release builds in the same workspace snapshot.

Table 6: Example Android package footprint (debug vs release, 2026-02-18)

Packaged item	Size	Notes
<code>example/build/app/outputs/flutter-apk/app-debug.apk</code>	178,454,872 bytes (≈ 170.19 MiB)	Example app debug APK.
<code>example/build/app/outputs/flutter-apk/app-release.apk</code>	113,030,559 bytes (≈ 107.80 MiB)	Example app release APK.
Release APK entry <code>lib/arm64-v8a/libkiwi.so</code>	7,613,192 bytes	Stripped native library entry inside APK.
Release APK entry <code>lib/x86_64/libkiwi.so</code>	11,381,344 bytes	Stripped native library entry inside APK.
Release APK model entries <code>assets/.../kiwi-models/cong/base/*</code>	79,574,759 bytes compressed	Same files are 99,308,057 bytes uncompressed.

Interpretation notes:

- Compressed and uncompressed measurements are not directly comparable.
- Source-tree native binaries and packaged APK entries represent different pipeline stages.
- Android packaging strips debug symbols from native libraries in this build flow, which dominates the source-vs-packaged size gap.
- ABI-specific native binaries should be interpreted per target split.
- Final store-delivery size can further differ by app-bundle splitting and distribution-side compression.

7 Public API Specification

7.1 Entry Point and Conditional Export

Public package entry point is `lib/flutter_kiwi_nlp.dart`. Runtime selection uses conditional exports:

```
export 'src/kiwi_analyzer_stub.dart'
  if (dart.library.io) 'src/kiwi_analyzer_native.dart'
  if (dart.library.js_interop) 'src/kiwi_analyzer_web.dart';
```

This guarantees a single import path for consumers while allowing runtime- specific implementation files.

7.2 Core Analyzer API

Table 7 lists the complete analyzer surface exposed to users.

Table 7: Core public analyzer API

Signature	Contract
<code>static Future<KiwiAnalyzer> create(...)</code>	Creates analyzer instance. Parameters include model path inputs, build/match flags, and thread count. Native resolves model path and opens FFI handle; web imports module and builds Kiwi instance (with fallback); stub throws unsupported exception.
<code>String get nativeVersion</code>	Returns backend version string. Native reads bridge-provided version. Web prefixes with <code>web/wasm</code> . Stub returns unsupported message.
<code>Future<KiwiAnalyzeResult> analyze(String text, {KiwiAnalyzeOptions options})</code>	Runs morphological analysis with candidate count and match options. Returns typed result object. Throws <code>KiwiException</code> on runtime failure or use-after-close.
<code>Future<void> addUserWord(String word, {String tag='NNP', double score=0.0})</code>	Registers user dictionary entry. Native invokes bridge function directly. Web stores word and rebuilds analyzer with accumulated <code>userWords</code> .
<code>Future<void> close()</code>	Releases resources and closes instance. Native closes bridge handle; web clears in-memory state; repeated use after close is rejected.

7.3 Options and Flags API

7.3.1 KiwiBuildOption constants

Table 8: Build option flags

Constant	Meaning
<code>integrateAllomorph</code>	Enable allomorph integration.
<code>loadDefaultDict</code>	Load default dictionary.
<code>loadTypoDict</code>	Load typo dictionary.
<code>loadMultiDict</code>	Load multi-word dictionary.
<code>modelTypeDefault</code>	Use backend default model type.
<code>modelTypeLargest</code>	Select largest model variant.

Constant	Meaning
<code>modelTypeKnlm</code>	Select KNLM model variant.
<code>modelTypeSbg</code>	Select SBG model variant.
<code>modelTypeCong</code>	Select CONG model variant.
<code>modelTypeCongGlobal</code>	Select global CONG variant.
<code>defaultOption</code>	Recommended bundle: <code>integrateAllomorph</code> <code>loadDefaultDict</code> <code>loadTypoDict</code> <code>loadMultiDict</code> <code>modelTypeCong</code> .

7.3.2 KiwiMatchOption constants

Table 9: Match option flags

Constant	Meaning
<code>url, email, hashtag, mention, serial</code>	Detect corresponding token classes.
<code>normalizeCoda</code>	Normalize final consonants.
<code>joinNounPrefix, joinNounSuffix, joinVerbSuffix, joinAdjSuffix, joinAdvSuffix</code>	Join morphology according to POS-specific rules.
<code>splitComplex</code>	Split complex forms.
<code>zCoda</code>	Enable coda-related matching behavior.
<code>compatibleJamo</code>	Emit compatibility jamo style output.
<code>splitSaisiot, mergeSaisiot</code>	Control sai-sios split/merge behavior.
<code>all</code>	Baseline option bundle.
<code>allWithNormalizing</code>	<code>all</code> plus <code>normalizeCoda</code> .

7.3.3 KiwiAnalyzeOptions

`KiwiAnalyzeOptions` carries request-level options: `topN` (candidate count) and `matchOptions` (bit-wise flags).

7.4 Result and Exception Models

Table 10: Public model types

Type	Fields and semantics
<code>KiwiToken</code>	<code>form</code> , <code>tag</code> , <code>offsets</code> (<code>start</code> , <code>length</code>), <code>positions</code> (<code>wordPosition</code> , <code>sentPosition</code>), confidence metrics (<code>score</code> , <code>typoCost</code>).
<code>KiwiCandidate</code>	<code>probability</code> and ordered <code>List<KiwiToken></code> sequence.
<code>KiwiAnalyzeResult</code>	<code>List<KiwiCandidate></code> candidates.
<code>KiwiException</code>	User-facing failure wrapper with message string; used for runtime, model, and lifecycle failures.

8 System Architecture Overview

Figure 4 summarizes how one Dart API fans out into native and web runtime lanes while preserving a shared contract.

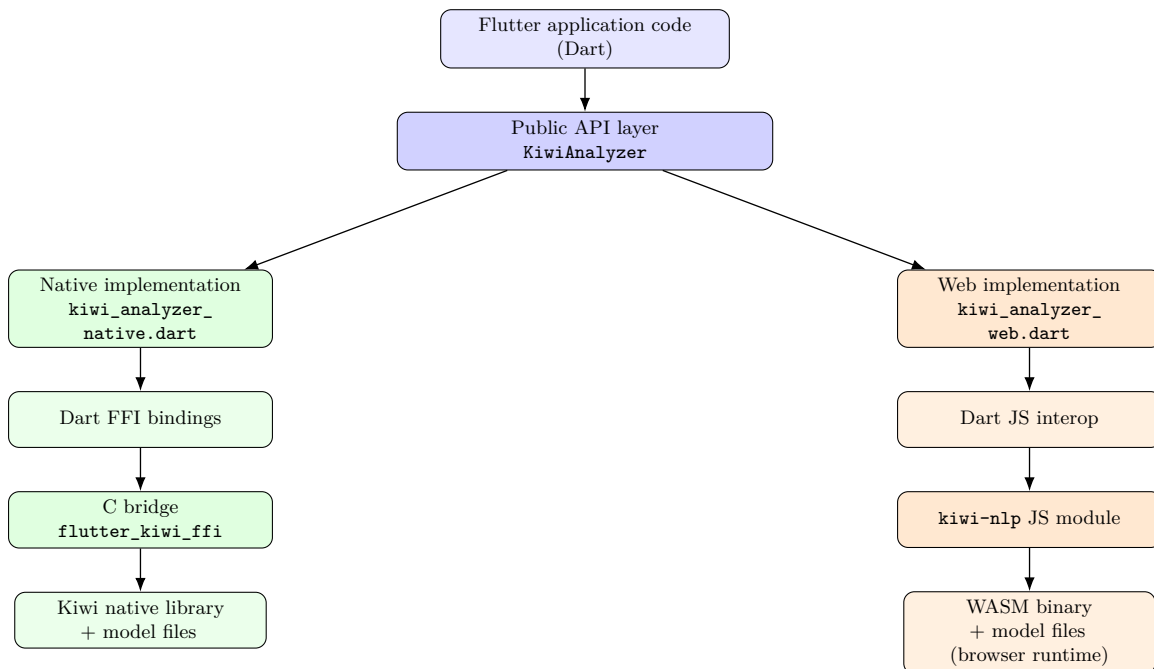


Figure 4: High-level plugin architecture for native and web paths.

8.1 Architecture Reading Notes

- The API entrypoint is intentionally singular (`KiwiAnalyzer`) to keep application code independent from target runtime.
- Native and web lanes diverge below the API surface, then converge back to the same typed Dart result models.
- Model artifacts remain external data assets in both lanes, but loading strategies differ by runtime constraints.

9 Native Runtime Implementation

9.1 Layered Architecture

Native execution is deliberately split into three layers:

1. Dart analyzer implementation (`lib/src/kiwi_analyzer_native.dart`).
2. C bridge shared library (`src/flutter_kiwi_ffi.c`, header `src/flutter_kiwi_ffi.h`).
3. Kiwi shared library (loaded dynamically by the bridge).

Dart calls generated bindings in `lib/flutter_kiwi_ffi_bindings_generated.dart`; the bridge owns Kiwi symbol resolution and native error buffering.

9.2 Bridge ABI

The bridge exports a compact ABI: `init`, `close`, `analyze_json`, `add_user_word`, `free_string`, `last_error`, and `version`.

The key design choice is JSON transport for analyze output. This decouples Dart from Kiwi internal structs and simplifies ownership management across language boundaries.

Initial JSON adoption was also a delivery-speed choice during early plugin bring-up. The analyzer output is a nested, variable-length schema (results/candidates/tokens), and a JSON boundary allowed stable cross-language representation before introducing a custom binary schema/allocator contract. In practical terms, the initial trade-off prioritized:

- schema evolution safety across bridge and Dart model changes,
- lower early risk for pointer-ownership bugs in nested payload paths,
- straightforward debugging with human-readable payloads.

The benchmark sections show the cost of that choice explicitly: measurable serialization overhead on hot paths, motivating binary/typed bridge evolution.

9.3 Dart FFI Binding Layer

The native path uses Dart's `dart:ffi` for symbol binding and `package:ffi/ffi.dart` for UTF-8/pointer utilities. At runtime:

1. `lib/src/kiwi_analyzer_native.dart` opens the bridge dynamic library.
2. `lib/flutter_kiwi_ffi_bindings_generated.dart` binds exported C symbols.
3. `GeneratedKiwiNativeBindings` adapts generated calls behind the `KiwiNativeBindings` interface for testability and swap-in fakes.

This structure keeps API code independent from raw pointer handling while still using zero-copy native handles for analyzer lifecycle management.

9.3.1 FFI Type/Ownership Mapping

- `flutter_kiwi_ffi_handle_t*` → opaque `Pointer<flutter_kiwi_ffi_handle_t>` in Dart.
- `int32_t` and `float` → Dart `int/double` via generated `asFunction()` bridges.
- `char*` inputs are allocated from Dart strings, passed as UTF-8, and explicitly released on the Dart side.
- `char*` outputs from `analyze_json()` are bridge-owned and must be released by `flutter_kiwi_ffi_free_string()`.
- Error text from `last_error()` is thread-local on the bridge side and consumed as read-only C strings from Dart.

9.4 ffigen Generation Pipeline

Binding code is generated (not handwritten) from the canonical header `src/flutter_kiwi_ffi.h`. The generation source-of-truth is `ffigen.yaml`, which declares:

- binding class name: `FlutterKiwifFiBindings`,
- entry/include header: `src/flutter_kiwi_ffi.h`,
- output file: `lib/flutter_kiwi_ffi_bindings_generated.dart`.

Regeneration command:

```
dart run ffigen --config ffigen.yaml
```

Why this paper uses `ffigen` instead of handwritten bindings:

- ABI drift resistance when C signatures evolve in `src/flutter_kiwi_ffi.h`.
- Deterministic regeneration from one header/config pair, which improves reviewability and incident forensics.
- Lower human error risk in repetitive signature plumbing (`NativeFunction`, `asFunction()`, pointer type mapping).
- Consistent symbol surface between C exports and Dart binding class.
- `FlutterKiwifFiBindings` keeps explicit bridge/API mapping and lowers long-term maintenance risk across Android/iOS/desktop targets.

9.4.1 Code-Generation Pipeline Diagram

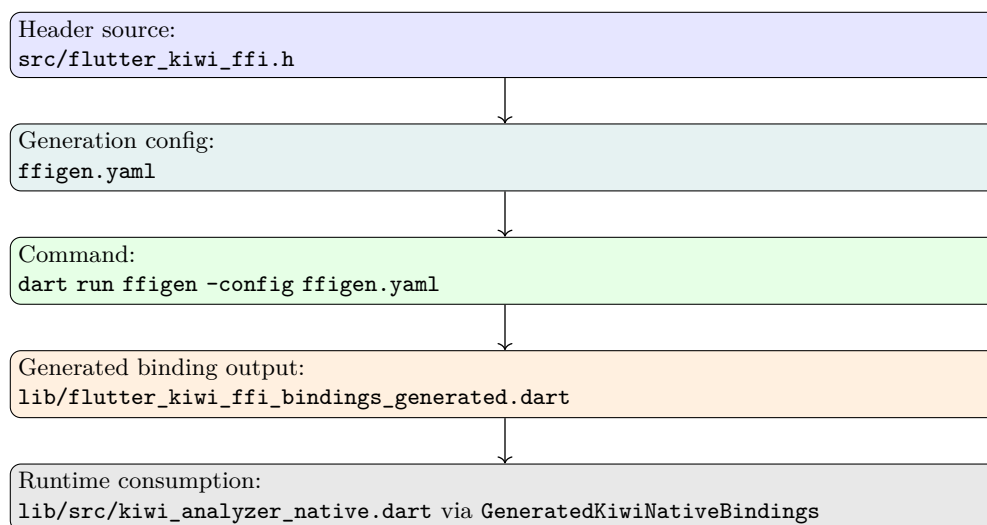


Figure 5: Dart FFI code-generation pipeline used in this plugin.

9.5 Dynamic Library Loading Strategy

9.5.1 Bridge loading (Dart side)

Candidate library names are tried in order by host platform:

- Apple: `flutter_kiwi_nlp.framework/flutter_kiwi_nlp`, then `flutter_kiwi_ffi.framework/flutter_kiwi_ffi`.
- Linux/Android: `libflutter_kiwi_ffi.so`, then `libflutter_kiwi_nlp.so`.
- Windows: `flutter_kiwi_ffi.dll`, then `flutter_kiwi_nlp.dll`.

9.5.2 Kiwi engine loading (C bridge side)

Bridge logic accepts optional override path via environment: `FLUTTER_KIWI_NLP_LIBRARY_PATH` (legacy alias: `FLUTTER_KIWI_FFI_LIBRARY_PATH`). Errors are captured in thread-local storage and exposed through `flutter_kiwi_ffi_last_error()`.

9.6 Analyzer Lifecycle and Memory Ownership

9.6.1 Creation

1. Resolve model directory using layered strategy.
2. Convert path to UTF-8 pointer.
3. Call `flutter_kiwi_ffi_init(...)`.
4. On null handle, read bridge error and throw `KiwiException`.
5. Free temporary path pointer.

9.6.2 Analysis

1. Ensure analyzer is open.
2. Convert input string to UTF-8 pointer.
3. Call `flutter_kiwi_ffi_analyze_json`.
4. Decode JSON into `KiwiAnalyzeResult`.
5. Release bridge-owned response string via `flutter_kiwi_ffi_free_string`.
6. Free input pointer.

9.6.3 Close

`close()` calls bridge close, marks instance closed, and rejects any subsequent operation with explicit use-after-close error messaging.

9.7 Native Model Resolution Algorithm

Native model resolution order is strict and deterministic:

1. `modelPath` argument.
2. `assetModelPath` argument.
3. Environment `FLUTTER_KIWI_NLP_MODEL_PATH` (legacy alias `FLUTTER_KIWI_FFI_MODEL_PATH`).
4. Compile-time define `FLUTTER_KIWI_NLP_ASSET_MODEL_PATH` (legacy alias `FLUTTER_KIWI_FFI_ASSET_MODEL_PATH`).
5. Built-in package asset candidates.
6. Download-and-cache fallback archive.

```
resolveModelPath(modelPath, assetModelPath):  
  if modelPath is non-empty: return modelPath  
  if assetModelPath is non-empty: return extractAssets(assetModelPath)  
  if env MODEL_PATH is non-empty: return env MODEL_PATH  
  if compile-time ASSET_MODEL_PATH is non-empty:  
    return extractAssets(ASSET_MODEL_PATH)  
  for candidate in builtInAssetCandidates:  
    if assetExists(candidate):  
      return extractAssets(candidate)  
  return ensureDownloadedModel()
```

9.8 Native Archive Fallback and Integrity

Fallback archive handling includes:

- configurable archive URL, cache key, and SHA-256 define,
- download timeout guard,
- partial-file strategy with atomic rename,
- extraction retry behavior,
- required-file completeness and minimum-size checks.

Required model files are currently: `combiningRule.txt`, `cong.mdl`, `default.dict`, `dialect.dict`, `extract.mdl`, `multi.dict`, `sj.morph`, and `typo.dict`.

In this plugin, that list is treated as strictly required at initialization time. The native model-file check validates all entries in `kiwiModelFileNames`, including `extract.mdl`, before analyzer construction proceeds.

10 Web Runtime Implementation

10.1 Module and WASM Bootstrap

Web runtime imports `kiwi-nlp` using configurable module and WASM URLs. Promise-like JS values are normalized through helper wrappers and transformed into Dart exceptions on rejection.

10.2 How WASM Runs on the Web in This Plugin

The browser execution path is a staged pipeline:

1. Flutter web code calls `KiwiAnalyzer.create()`.
2. Dart JS interop loads the configured `kiwi-nlp` JavaScript module URL.
3. The JS module resolves and instantiates its WASM binary from `FLUTTER_KIWI_NLP_WEB_WASM_URL` (or module defaults).
4. Model files are loaded by URL map when available; otherwise archive fallback is triggered and extracted in memory.
5. A Kiwi runtime instance is constructed in browser memory and referenced by the Dart wrapper.
6. Each `analyze()` call crosses the Dart-to-JS boundary, executes in WASM-backed logic, then returns JS data converted into typed Dart models.

Important constraint: this plugin does not assume a browser-available native ABI or direct file-system model layout. All web model handling is network and in-memory oriented by design.

10.3 Web Bootstrap Latency and TTI Considerations

Web UX is affected not only by steady-state throughput but also by initial bootstrap latency. In this plugin, a first-use web interaction latency can be conceptually decomposed as:

$$T_{\text{TTI}} \approx T_{\text{module fetch}} + T_{\text{wasm fetch}} + T_{\text{model fetch}} + T_{\text{instantiate}} + T_{\text{first analyze}}.$$

Current report scope focuses on native/mobile throughput and correctness; it does not yet include sustained web TTI measurements under controlled cache and network conditions. This is treated as explicit future work in Section 18.3. For reproducible Web TTI reporting in the next revision, we will publish cold-cache and warm-cache TTI distributions (p50/p95) under explicit network profiles, together with module/WASM/model artifact sizes and cache-hit/miss annotations.

10.4 Web Build Modes

The web implementation supports two construction modes:

1. **Direct builder mode:** `KiwiBuilder.build()` returns a Kiwi object.
2. **API bridge mode:** when builder API is available, model files are loaded through API methods and analyzer commands are sent by `cmd(...)` with Kiwi instance id.

This dual-path logic reduces fragility against upstream module behavior changes.

10.5 Web Fallback Algorithm

If URL-based model loading fails, the runtime attempts archive fallback:

1. try explicit archive URL define,
2. try default release URL composed from repo/version/name,

3. try GitHub releases API metadata lookup,
4. download bytes, optional SHA-256 verification,
5. extract required files in memory,
6. validate completeness/minimum size,
7. retry Kiwi build with in-memory model map.

10.6 Web-Specific Behavioral Notes

- `numThreads` and `matchOptions` are accepted in `create()` for API parity but are not applied at web create phase.
- `addUserWord()` triggers a rebuild with accumulated `userWords` rather than direct mutable insertion.
- `nativeVersion` reports `web/wasm <version>` format.

11 Build and Packaging Pipeline

11.1 Android

- Gradle task `prepareKiwiAndroidLibs` runs before `preBuild`.
- Default ABIs: `arm64-v8a`, `x86_64`.
- Script: `tool/build_android_libkiwi.sh`.
- Output: `android/src/main/jniLibs/<abi>/libkiwi.so`.

11.2 iOS

- Hook: `ios/flutter_kiwi_nlp.podspec` prepare command.
- Script: `tool/build_ios_kiwi_xcframework.sh`.
- Output: `ios/Frameworks/Kiwi.xcframework`.
- Includes device and simulator slices in one `XCFramework`.

11.3 macOS

- Hook: `macos/flutter_kiwi_nlp.podspec` prepare command.
- Script: `tool/build_macos_kiwi_dylib.sh`.
- Output: `macos/Frameworks/libkiwi.dylib`.
- Supports `arm64/x86_64` merge via `lipo`.

11.4 Linux

- Hook: `linux/CMakeLists.txt` custom target `prepare_kiwi_linux_lib`.
- Script: `tool/build_linux_libkiwi.sh`.
- Output: `linux/prebuilt/libkiwi.so`.
- Strategy: prebuilt download first, source build fallback.

11.5 Windows

- Hook: `windows/CMakeLists.txt` custom target `prepare_kiwi_windows_dll`.
- Script: `tool/build_windows_kiwi_dll.ps1`.
- Output: `windows/prebuilt/kiwi.dll`.
- Strategy: prebuilt download first (where available), source build fallback.

11.6 Shared Bridge Build

`src/CMakeLists.txt` builds shared bridge library `flutter_kiwi_ffi`. On Android, linker option `-Wl,-z,max-page-size=16384` is applied for Android 15 page-size compatibility; Linux links `dl` where required.

12 Configuration Surface

12.1 Native Configuration Keys

Table 11: Native runtime configuration keys

Key	Role
<code>FLUTTER_KIWI_NLP_MODEL_PATH</code>	Runtime environment model directory override.
<code>FLUTTER_KIWI_NLP_ASSET_MODEL_PATH</code>	Compile-time default packaged asset base path.
<code>FLUTTER_KIWI_NLP_MODEL_ARCHIVE_URL</code>	Compile-time archive URL for fallback download.
<code>FLUTTER_KIWI_NLP_MODEL_ARCHIVE_SHA256</code>	Optional checksum for archive integrity validation.
<code>FLUTTER_KIWI_NLP_MODEL_CACHE_KEY</code>	Cache directory discriminator for extracted archive assets.
<code>FLUTTER_KIWI_NLP_LIBRARY_PATH</code>	Runtime override for Kiwi shared library location.

Legacy aliases prefixed with `FLUTTER_KIWI_FFI_...` are retained for backward compatibility.

12.2 Web Configuration Keys

Table 12: Web runtime configuration keys

Key	Role
FLUTTER_KIWI_NLP_WEB_MODULE_URL	JavaScript module URL for <code>kiwi-nlp</code> .
FLUTTER_KIWI_NLP_WEB_WASM_URL	WASM binary URL passed to builder.
FLUTTER_KIWI_NLP_WEB_MODEL_BASE_URL	Base path used to construct per-file model URL map.
FLUTTER_KIWI_NLP_WEB_MODEL_ARCHIVE_URL	Optional explicit archive URL for fallback model download.
FLUTTER_KIWI_NLP_WEB_MODEL_ARCHIVE_SHA256	Optional archive checksum verification value.
FLUTTER_KIWI_NLP_WEB_MODEL_GITHUB_REPO	Repository slug used for release metadata fallback lookup.
FLUTTER_KIWI_NLP_WEB_MODEL_ARCHIVE_VERSION	Release tag used when composing default archive URL.
FLUTTER_KIWI_NLP_WEB_MODEL_ARCHIVE_NAME	Archive filename used in default/fallback URL generation.

13 Failure Taxonomy and Mitigations

Table 13: Failure categories and implemented mitigations

ID	Failure condition	Mitigation strategy
F1	Bridge library not loadable	Multi-candidate load attempt plus aggregated error diagnostics.
F2	Kiwi library unresolved or symbols missing	Runtime path override support and explicit bridge last-error propagation.
F3	Model path unresolved	Ordered fallback chain (arguments, env, defines, assets, archive).
F4	Archive download failure (timeout/HTTP)	Retry path and fallback URL sequence; actionable exception text.
F5	Archive integrity/completeness failure	SHA-256 verification option and minimum-size checks per required model file.
F6	Web module import/promise rejection	Promise normalization wrappers, timeout guards, and error contextualization.
F7	API use after close	Explicit lifecycle state check and deterministic <code>KiwiException</code> .
F8	Native library crash (for example segmentation fault in dependent native code)	No in-process isolation in current design; such crashes are fatal to the host Flutter process and must be mitigated by upstream native stability and artifact validation.

14 Performance Evaluation and Benchmark Interpretation

14.1 Evaluation Objective

The benchmark goal is to compare end-to-end analyzer behavior between `flutter_kiwi_nlp` and `kiwipiepy` under a shared corpus and roughly aligned loop structure. The result is intended as an engineering signal, not as a definitive language-model quality ranking.

14.2 Implemented Benchmark Pipeline

The repository executes comparison in three stages:

1. `tool/benchmark/run_compare.py` launches Flutter benchmark app (target: `example/lib/benchmark_main.dart`) and captures JSON payload from benchmark marker lines in stdout/device logs.
2. The same runner executes `tool/benchmark/kiwipiepy_benchmark.py` on Python runtime.
3. `tool/benchmark/compare_results.py` merges both JSON files into one markdown table with mean/stddev and per-trial raw snapshots.

Canonical corpus file: `example/assets/benchmark_corpus_ko.txt`. The runner supports repeated independent trials through `-trials`, producing both per-trial JSON files and aggregated report.

14.3 Reproducibility Manifest

Table 14: Execution metadata used for the benchmark section

Item	Value
Report snapshot date	February 18, 2026
Repository commit base	2ab94d2b3fa6a941c8a339547e4e1797695ab06d
Flutter SDK	3.41.1 (framework revision 582a0e7c55)
Dart SDK	3.11.0
Python runtime	3.14.3
kiwipiepy	0.22.2
Xcode toolchain	Xcode 26.2 (17C52)
Desktop host OS for baseline table	macOS-15.7.4-arm64-arm-64bit-Mach-0
Host CPU	Apple M2 Pro, 10 logical cores
Host memory	16 GiB (17179869184 bytes)
Corpus hash (SHA-256)	0fed28f4601cd577de8ad0f35fbe5bb1827e71931d3c8b19714a9976a12f3c9f
Desktop benchmark shape	trials=3, warmup_runs=3, measure_runs=15, top_n=1
Desktop analyze impl	Flutter: json (primary), Kiwi: analyze
Linux/Windows CI runner shape	trials=3, warmup_runs=3, measure_runs=15, top_n=1

Item	Value
Linux/Windows CI analyze impl	Flutter: <code>token_count</code> (primary), Kiwi: <code>analyze</code>

14.4 Recorded Configuration and Workload

From the benchmark trial sets included in this report (February 18, 2026):

- Desktop reference platform: macOS 15.7.4 arm64 (release, `n=3`, `warmup=3`, `measure=15`).
- Mobile platform A: iOS 26.2 simulator (iPhone 17, debug, `n=3`, `warmup=3`, `measure=15`).
- Mobile platform B: Android 16 emulator (API 36, release, `n=3`, `warmup=3`, `measure=15`).
- Desktop CI platform A: Linux GitHub runner (`n=3`, `warmup=3`, `measure=15`).
- Desktop CI platform B: Windows GitHub runner (`n=3`, `warmup=3`, `measure=15`).
- Corpus sentences: 40.
- Sample POS rows emitted per runtime: `sample_count` = 10.
- Total measured analyses per trial: all reported rows use $15 * 40 = 600$.
- `top_n`: 1.
- Desktop primary analyze path: Flutter `json`, Kiwi `analyze` (API path differs; see artifact caution note).
- Build options: 1039 (`integrateAllomorph`, `loadDefaultDict`, `loadTypoDict`, `loadMultiDict`, `modelTypeCong`).
- Analyze match options: 8454175 (`allWithNormalizing` bundle).
- Flutter analyzer setting: `numThreads` = -1.
- Python analyzer setting: `num_workers` = -1.
- iOS Simulator did not support `release/profile` in this setup, so iOS measurements were collected in `debug` mode.
- For mobile runs, `kiwipiepy` comparison values were collected on the host macOS runtime using the same corpus and benchmark loop settings.
- This revision introduces layered timing decomposition fields: `pure_elapsed_ms`, `full_elapsed_ms`, and `json_overhead_ms`, in addition to warm/cold summaries.

Item	Value
------	-------

14.5 Mobile Test Environment Details

Table 15: iOS/Android benchmark environment details

Item	Value
iOS test target	iPhone 17 simulator, UDID <REDACTED_SIM_UDID>
iOS runtime	com.apple.CoreSimulator.SimRuntime.iOS-26-2, iOS 26.2 (build 23C54), supported architecture arm64
iOS device type	com.apple.CoreSimulator.SimDeviceType.iPhone-17, model identifier iPhone18,3
iOS CPU/memory context	simctl spawn sysctl: hw.ncpu=10, hw.memsize=17179869184. In this setup, simulator-reported values match host resources.
Android test target	<REDACTED_EMULATOR_ID>, AVD name Pixel_9, guest model sdk_gphone64_arm64
Android guest OS/ABI	Android 16 (ro.build.version.sdk=36), ABI arm64-v8a, hardware ranchu
Android virtual hardware config	AVD config.ini: hw.cpu.ncore=4, hw.ramSize=2048, resolution 1080x2424, density 420, data partition 6G
Android runtime observation	/proc/meminfo: MemTotal=2017772 kB; /proc/cpuinfo: 4 processors visible in guest

Environment data was captured immediately after benchmark runs using `flutter devices`, `sysctl`, `xcrun simctl`, `adb`, and AVD `config.ini` inspection.

14.6 Metric Definitions

Both benchmark programs produce comparable derived metrics: **Cold start** in this paper means one-time analyzer initialization from a not-ready state (library/model load and analyzer construction). **Warm path** means steady-state `analyze()` execution after initialization has already completed.

These are reported separately because they answer different operational questions: cold start dominates short sessions and first-use latency, while warm path dominates sustained throughput/latency under repeated analysis calls. **Session-length effective throughput** then combines both to show end-to-end impact for finite request counts.

- warm metrics use only measured loop time (`elapsed_ms`) after warmup; cold init is excluded from throughput/latency rows.
- `init (ms)` = elapsed wall time around analyzer creation call (`KiwiAnalyzer.create(...)`), including model-path resolution, potential asset extraction/copy from Flutter bundle to local temp path, and native bridge/Kiwi initialization.
- `analyses/sec` = `total_analyses` / `elapsed_seconds`
- `chars/sec` = `total_chars` / `elapsed_seconds`

- $\text{tokens/sec} = \text{total_tokens} / \text{elapsed_seconds}$
- $\text{avg latency (ms)} = \text{elapsed_ms} / \text{total_analyses}$
- $\text{avg token latency (us/token)} = \text{elapsed_ms} * 1000 / \text{total_tokens}$
- session-length effective analyses/sec (init included) for N analyses:

$$\text{effective_analyses_per_sec} = \frac{N}{(\text{init_ms}/1000) + (N/\text{analyses_per_sec})}.$$

Using notation aligned with benchmark payload fields:

$$S = \frac{\text{elapsed_ms}}{1000}, \quad A = \text{total_analyses}, \quad C = \text{total_chars}, \quad U = \text{total_tokens}.$$

$$T_{\text{analyses}} = \frac{A}{S}, \quad T_{\text{chars}} = \frac{C}{S}, \quad T_{\text{tokens}} = \frac{U}{S}.$$

$$L_{\text{avg-ms}} = \frac{\text{elapsed_ms}}{A}, \quad L_{\text{token-us}} = \frac{1000 \cdot \text{elapsed_ms}}{U}.$$

$$T_{\text{eff}}(N) = \frac{N}{(\text{init_ms}/1000) + (N/T_{\text{analyses}})}.$$

14.7 Boundary-Decomposed Measurement (Pure vs Full Path)

To make performance interpretation more granular, the updated benchmark captures both:

- **Pure path:** analyzer compute path using `token_count`-based timing.
- **Full path:** full JSON analyze payload path (`json path`) including bridge serialization/parsing overhead.

Let E_{pure} and E_{full} denote elapsed times for pure/full paths on the same trial workload. Then:

$$\text{BoundaryLoss} = 1 - \frac{E_{\text{pure}}}{E_{\text{full}}} = 1 - \frac{T_{\text{full}}}{T_{\text{pure}}}.$$

Table 16: Layered boundary decomposition summary (mean \pm stddev)

Env.	Pure APS	Full APS	Loss	Ovh./analysis (ms)	Ovh. ratio
macOS desktop baseline	12044.53 \pm 168.39	7606.26 \pm 276.63	36.85%	0.0485 \pm 0.0046	36.85 \pm 2.27%
iOS simulator (debug)	9890.53 \pm 3756.49	5150.22 \pm 1424.86	47.93%	0.0891 \pm 0.0729	41.81 \pm 25.87%
Android emulator (release)	3865.71 \pm 811.36	3094.32 \pm 930.90	19.95%	0.0810 \pm 0.0630	21.15 \pm 8.57%

14.8 Recorded Result Summary: Desktop Baseline (macOS, n=3)

Table 17: macOS desktop warm-path summary (mean \pm stddev)

Metric	flutter_kiwi_nlp	kiwipiepy	Ratio (Flutter/Kiwi)
Throughput (analyses/s)	7606.26 \pm 276.63	2439.44 \pm 29.54	3.12x faster
Throughput (chars/s)	257852.20 \pm 9377.59	82697.06 \pm 1001.31	3.12x faster
Throughput (tokens/s)	122460.78 \pm 4453.66	39214.02 \pm 474.81	3.12x faster
Avg latency (ms, lower better)	0.13 \pm 0.00	0.41 \pm 0.00	0.32x faster
Avg token latency (us/token, lower better)	8.17 \pm 0.29	25.50 \pm 0.31	0.32x faster

14.9 Statistical Interval View: Desktop Warm Path (95% CI)

To make variability interpretation more explicit, Table 18 adds 95% confidence intervals for desktop baseline means ($n = 3$, two-sided t -interval, $df = 2$).

$$CI_{95}(\mu) = \bar{x} \pm t_{0.975, n-1} \cdot \frac{s}{\sqrt{n}}.$$

Here s denotes the sample standard deviation across trials and $t_{0.975, n-1}$ is the Student- t quantile for two-sided 95% intervals. This interval assumes independent repeated trials under the same benchmark configuration.

Table 18: Desktop baseline 95% confidence intervals

Metric	flutter_kiwi_nlp (95% CI)	kiwipiepy (95% CI)
Throughput (analyses/s)	[6919.07, 8293.45]	[2366.06, 2512.82]
Throughput (chars/s)	[234556.98, 281147.42]	[80209.67, 85184.45]
Throughput (tokens/s)	[111397.28, 133524.28]	[38034.53, 40393.51]
Avg latency (ms)	[0.130, 0.130]	[0.410, 0.410]
Avg token latency (us/token)	[7.45, 8.89]	[24.73, 26.27]

14.10 Cold-Start Summary (Median + p95)

Table 19: Cold-start init summary reported separately from warm metrics

Environment	flutter_kiwi_nlp init (ms)	kiwipiepy init (ms)	Ratio (median)
macOS desktop baseline	median 1669.67, p95 1675.29	median 894.68, p95 914.06	1.87x slower
iOS simulator (debug)	median 1740.17, p95 1746.64	median 848.42, p95 885.88	2.05x slower
Android emulator (release)	median 4114.19, p95 4273.30	median 1020.01, p95 1140.11	4.03x slower

Interpretation note: this cold-start metric is intentionally end-to-end and not micro-phased. On the Flutter path, it may include asset-model extraction/copy work on cache-miss paths in addition

to native library/bridge initialization. Therefore, the Android 4.03x median gap should be read as combined startup cost, not as a JNI/FFI-only penalty estimate.

14.11 Additional Mobile Runtime Measurements (iOS n=3, Android n=3)

14.11.1 iOS Simulator (debug)

Table 20: iOS simulator warm-path summary (mean \pm stddev)

Metric	flutter_kiwi_nlp	kiwipiepy	Ratio (Flutter/Kiwi)
Throughput (analyses/s)	9890.53 \pm 3756.49	2553.35 \pm 61.83	3.87x faster
Throughput (chars/s)	335288.96 \pm 127345.03	86558.61 \pm 2096.03	3.87x faster
Throughput (tokens/s)	159237.53 \pm 60479.50	41045.12 \pm 993.92	3.88x faster
Avg latency (ms, lower better)	0.11 \pm 0.05	0.39 \pm 0.01	0.29x faster
Avg token latency (us/token, lower better)	7.11 \pm 3.30	24.37 \pm 0.58	0.29x faster

14.11.2 Android Emulator (release)

Table 21: Android emulator warm-path summary (mean \pm stddev)

Metric	flutter_kiwi_nlp	kiwipiepy	Ratio (Flutter/Kiwi)
Throughput (analyses/s)	3865.71 \pm 811.36	2543.80 \pm 69.61	1.52x faster
Throughput (chars/s)	131047.62 \pm 27505.16	86234.83 \pm 2359.65	1.52x faster
Throughput (tokens/s)	62237.96 \pm 13062.92	40891.59 \pm 1118.92	1.52x faster
Avg latency (ms, lower better)	0.27 \pm 0.06	0.39 \pm 0.01	0.68x faster
Avg token latency (us/token, lower better)	16.62 \pm 3.96	24.47 \pm 0.68	0.68x faster

14.12 Additional Desktop CI Measurements (Linux/Windows, n=3)

Table 22: Linux/Windows GitHub runner warm-path summary (single vs batch, mean \pm stddev)

Platform/Mode	Flutter APS	Kiwi APS	Ratio	Flutter Lat (ms)	Kiwi Lat (ms)	Ratio
Linux/single	2084.75 \pm 74.15	3244.11 \pm 48.99	0.64x	0.48 \pm 0.02	0.31 \pm 0.00	1.56x slower
Linux/batch	5129.13 \pm 56.53	3321.86 \pm 6.59	1.54x	0.19 \pm 0.00	0.30 \pm 0.00	0.65x faster
Windows/single	1921.24 \pm 5.33	3018.30 \pm 67.70	0.64x	0.52 \pm 0.00	0.33 \pm 0.01	1.57x slower

Platform/Mode	Flutter APS	Kiwi APS	Ratio	Flutter Lat (ms)	Kiwi Lat (ms)	Ratio
Windows/batch	4622.39 \pm 35.36	2987.51 \pm 23.77	1.55x	0.22 \pm 0.00	0.33 \pm 0.00	0.65x faster

Table 23: Linux/Windows GitHub runner boundary/cold/session summary

Platform/Mode	Boundary Loss	Ovh./analysis (ms)	Init Ratio (median)	N=1000 Eff. Ratio
Linux/single	13.89%	0.0779 \pm 0.0114	2.69x slower	0.42x
Linux/batch	22.22%	0.0557 \pm 0.0025	2.71x slower	0.49x
Windows/single	15.85%	0.0982 \pm 0.0094	2.67x slower	0.42x
Windows/batch	23.12%	0.0651 \pm 0.0054	2.54x slower	0.47x

14.13 Session-Length Effective Throughput (Init Included)

Table 24 reports expected analyses/sec when a session runs for fixed analysis counts (N in {1, 10, 100, 1000}), combining cold init and warm throughput in one user-facing rate.

Table 24: Session-length effective throughput (mean \pm stddev)

Environment	N	flutter_kiwi_nlp	kiwipiepy	Ratio (Flutter/Kiwi)
macOS desktop baseline	1	0.60 \pm 0.01	1.11 \pm 0.02	0.54x
macOS desktop baseline	10	6.01 \pm 0.06	11.10 \pm 0.22	0.54x
macOS desktop baseline	100	59.65 \pm 0.60	106.60 \pm 1.97	0.56x
macOS desktop baseline	1000	557.18 \pm 5.11	765.03 \pm 8.08	0.73x
iOS simulator (debug)	1	0.58 \pm 0.00	1.16 \pm 0.03	0.50x
iOS simulator (debug)	10	5.75 \pm 0.04	11.56 \pm 0.32	0.50x
iOS simulator (debug)	100	57.17 \pm 0.42	111.05 \pm 2.87	0.51x
iOS simulator (debug)	1000	540.24 \pm 15.79	797.83 \pm 9.76	0.68x
Android emulator (release)	1	0.24 \pm 0.01	0.96 \pm 0.09	0.25x
Android emulator (release)	10	2.42 \pm 0.08	9.59 \pm 0.87	0.25x
Android emulator (release)	100	24.03 \pm 0.79	92.73 \pm 8.19	0.26x
Android emulator (release)	1000	227.05 \pm 5.21	697.44 \pm 46.98	0.33x

14.14 Visual Summary Charts (Desktop Baseline)

To improve reviewer readability, this paper includes normalized benchmark charts. Figures 6 and 7 use Kiwi mean as baseline index 100.

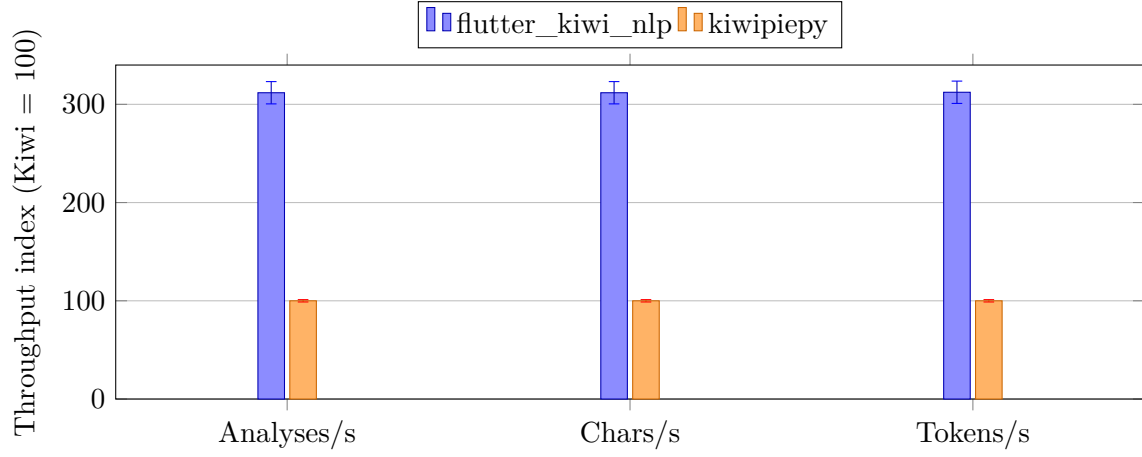


Figure 6: Throughput comparison for macOS baseline (normalized index, mean \pm stddev).

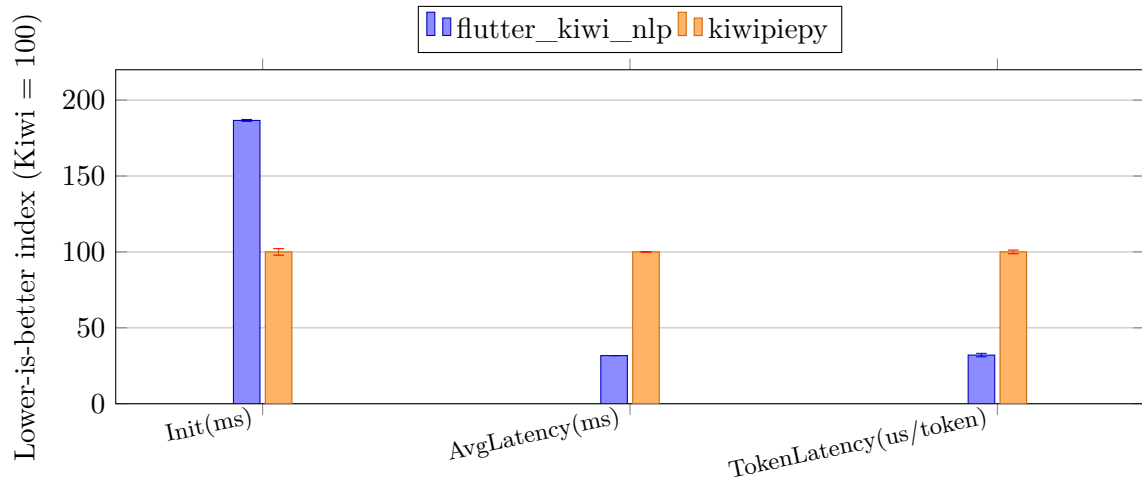


Figure 7: Lower-is-better metrics for macOS baseline (normalized index, mean \pm stddev).

14.15 Per-Trial Raw Trace (Desktop Baseline)

Table 25: Per-trial raw snapshot for macOS baseline traceability

Trial	Flutter init (ms)	Kiwi init (ms)	Flutter analyses/s	Kiwi analyses/s
1	1644.20	916.21	7478.97	2473.21
2	1675.92	894.68	7416.20	2418.41
3	1669.67	880.94	7923.62	2426.70

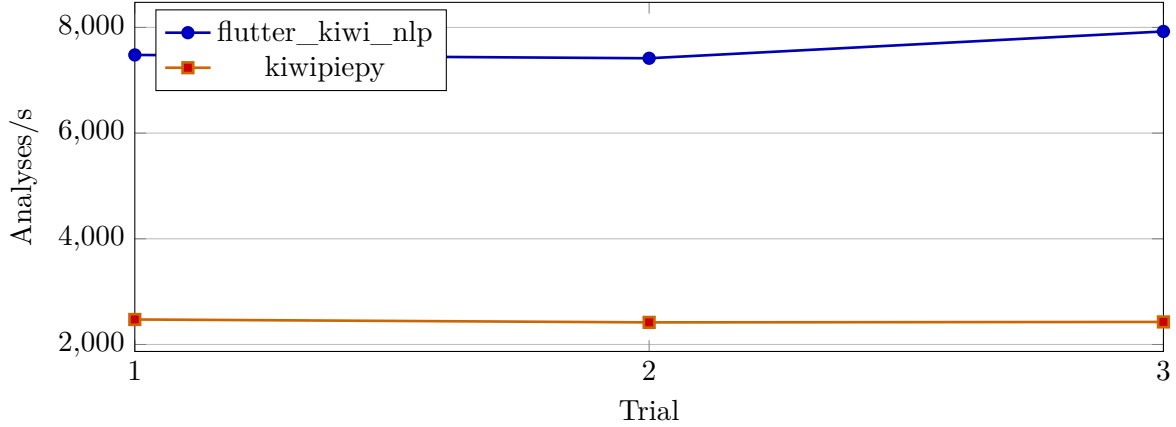


Figure 8: Trial-by-trial throughput trace for macOS baseline (analyses per second).

14.16 Detailed Interpretation (Honest Reading)

Across desktop and mobile trial sets, warm-path throughput and latency are consistently stable, and warm per-call latency remains sub-millisecond in these configurations (roughly 0.1–0.6 ms). This matches the practical impression that repeated in-app analysis is responsive after initialization. At the same time, repeated measurements still show slower init on the Flutter path relative to the Python reference runtime, with magnitude varying by environment. In the latest run profile, warm-path throughput itself reverses direction and is higher on Flutter across the macOS+iOS+Android baseline rows: desktop 3.12 \times , iOS simulator 3.87 \times , and Android emulator 1.52 \times versus `kiwipiepy`. Compared with the prior table values in this report, Flutter warm analyses/s increased sharply (desktop: 2546.24 \rightarrow 7606.26, +198.73%; iOS: 2422.31 \rightarrow 9890.53, +308.32%; Android: 2214.79 \rightarrow 3865.71, +74.54%). However, session-length effective throughput remains lower for short sessions because cold-start cost is still front-loaded on the Flutter path.

New desktop CI rows on Linux/Windows add an additional signal: execution mode is a first-order factor. In both OSes, single mode shows lower warm throughput than Python ($\approx 0.64\times$), while batch mode flips the direction and exceeds Python ($\approx 1.54\text{--}1.55\times$). This pattern is consistent with reduced per-call boundary overhead under batch dispatch, while init-heavy short sessions remain disadvantaged (Linux/Windows $N = 1000$ effective ratio still around 0.42–0.49 \times).

New boundary-decomposed instrumentation (Tables 16 and 23) adds direct evidence for where overhead accumulates. Flutter boundary loss (full JSON path vs pure processing path) spans 13.89–47.93% across the measured rows, with per-analysis overhead roughly 0.05–0.10 ms. This also bounds expected gains from JSON-path removal alone: eliminating serialization would likely recover a meaningful fraction of warm-path overhead, but not necessarily close the entire cross-runtime gap by itself.

Based on code inspection and payload values, the most likely contributors are:

1. **Boundary conversion overhead (evidence-based):** native Flutter path returns JSON from

C bridge (`flutter_kiwi_ffi_analyze_json`) and then parses it in Dart (`jsonDecode`). Python path consumes native results via a different binding stack without this exact JSON roundtrip. In bridge code, JSON is assembled through repeated `sb_append(...)` operations and escaped field appends (`sb_append_json_escaped`), which implies dynamic allocation/reallocation and additional $O(n)$ -scale string-copy work on top of base inference.

2. **Per-call async boundary cost (evidence-based):** Flutter benchmark loop awaits `Future` analysis call for each sentence, adding runtime overhead at Dart async/FFI boundaries.
3. **Residual cross-runtime semantic gap (evidence-based):** this benchmark now passes explicit build and match bitmasks to both runtimes, reducing prior configuration mismatch risk. However, output token totals still differ in the current macOS baseline row (Flutter 9660 vs Python 9645), indicating backend-wrapper semantic differences even under aligned knobs.
4. **Worker/thread auto mode mismatch (evidence-based):** both runtimes use auto settings (-1) for thread/worker counts, but auto-mode policies are not guaranteed to map to equivalent parallel execution strategy.
5. **Model loading path variability (inference):** unless `-model-path` is forced, each runtime may initialize using different default lookup paths/caches, potentially affecting cold-start cost.

14.17 Threats to Validity

Current benchmark limitations that should be stated explicitly:

- Trial counts are improved but still modest for strong inference (all reported rows use `n=3`).
- No explicit CPU pinning or thermal-state control.
- Linux/Windows desktop CI rows were collected on shared GitHub-hosted runners, so VM scheduling variability can affect jitter and tail latency.
- Bitmask parity is enforced, but internal backend semantics can still differ.
- Init metric combines multiple concerns (library load, model resolution, analyzer construction) rather than isolated micro-phases.
- iOS measurements in this report were collected on simulator in `debug` mode, because `release/profile` was unavailable in the current simulator setup.
- iOS simulator reports host-shared CPU/memory context (10 cores, 16 GiB RAM).
- These values are not equivalent to physical-device thermal or power constraints.
- Android measurements were collected on emulator, not on real devices.
- Android emulator guest was configured with 4 vCPUs and 2 GiB RAM, which can amplify startup variance and may not represent flagship physical devices.
- Physical-device benchmark commands are provided in Appendix C, but this revision does not yet include executed iOS/Android physical-device rows.
- For mobile rows, `kiwipiepy` values come from host macOS runtime; therefore cross-runtime ratios on mobile rows should be interpreted as engineering reference, not strict same-device head-to-head evidence.

- Web runtime functionality is validated in tests, but sustained web throughput benchmark numbers are not included in this report.
- Gold-corpus quality evaluation in this revision covers two Kiwi-provided evaluation sets (191 sentences total). Broader domain coverage still requires larger and more heterogeneous corpora.

14.18 Gold-Corpus Linguistic Agreement Evaluation

To close the prior quality-evidence gap, this revision adds a reproducible gold-corpus evaluation using:

- `example/assets/gold_eval_web_ko.txt` (158 sentences),
- `example/assets/gold_eval_written_ko.txt` (33 sentences),
- shared options for both runtimes: `top_n=1` and `build_options=1039`.
- match options: `create_match_options=8454175` and `analyze_match_options=8454175`.

Evaluation command:

```
uv run python tool/benchmark/gold_corpus_compare.py \
--device macos --mode release
```

Metrics are sequence-level agreements based on normalized Levenshtein distance: token agreement compares **form** sequences, POS agreement compares **form/tag** sequences, and sentence exact match requires full sequence identity. Let M be the number of evaluated sentences. For sentence i , g_i^{form} and p_i^{form} are gold/predicted token-form sequences, and g_i^{pair} , p_i^{pair} are gold/predicted **form/tag** pair sequences.

$$A_{\text{token}} = 1 - \frac{\sum_{i=1}^M d_{\text{Lev}}(g_i^{\text{form}}, p_i^{\text{form}})}{\sum_{i=1}^M \max(|g_i^{\text{form}}|, |p_i^{\text{form}}|)}.$$

$$A_{\text{pos}} = 1 - \frac{\sum_{i=1}^M d_{\text{Lev}}(g_i^{\text{pair}}, p_i^{\text{pair}})}{\sum_{i=1}^M \max(|g_i^{\text{pair}}|, |p_i^{\text{pair}}|)}.$$

$$E_{\text{token-seq}} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}[g_i^{\text{form}} = p_i^{\text{form}}], \quad E_{\text{sentence}} = \frac{1}{M} \sum_{i=1}^M \mathbf{1}[g_i^{\text{pair}} = p_i^{\text{pair}}].$$

Table 26: Gold-corpus overall agreement (191 sentences, 7,990 gold tokens)

Metric	flutter_kiwi_nlp	kiwipiepy	Delta (pp)
Token agreement	88.39%	88.58%	-0.19
POS agreement	84.90%	85.55%	-0.65
Token-sequence exact match	3.66%	3.66%	+0.00
Sentence exact match	1.57%	2.62%	-1.05

Table 27: Per-dataset agreement breakdown

Dataset	Runtime	Token (%)	POS (%)	Token exact (%)	Sentence exact (%)
web_ko	flutter_kiwi_nlp	87.79	84.04	3.80	1.90
web_ko	kiwipiepy	88.03	84.80	3.80	2.53
written_ko	flutter_kiwi_nlp	90.86	88.42	3.03	0.00
written_ko	kiwipiepy	90.86	88.61	3.03	3.03

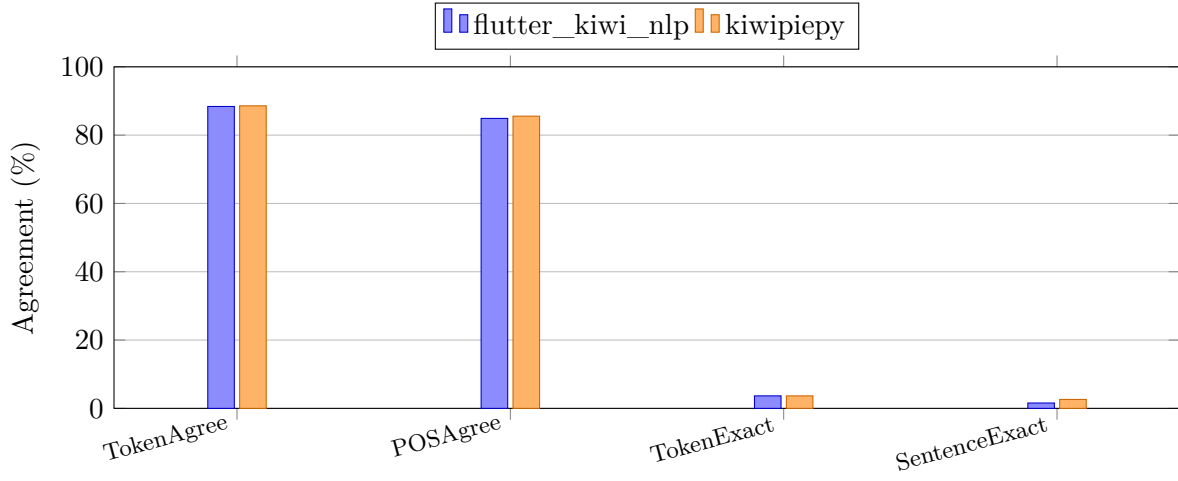


Figure 9: Overall gold-corpus agreement metrics across runtimes.

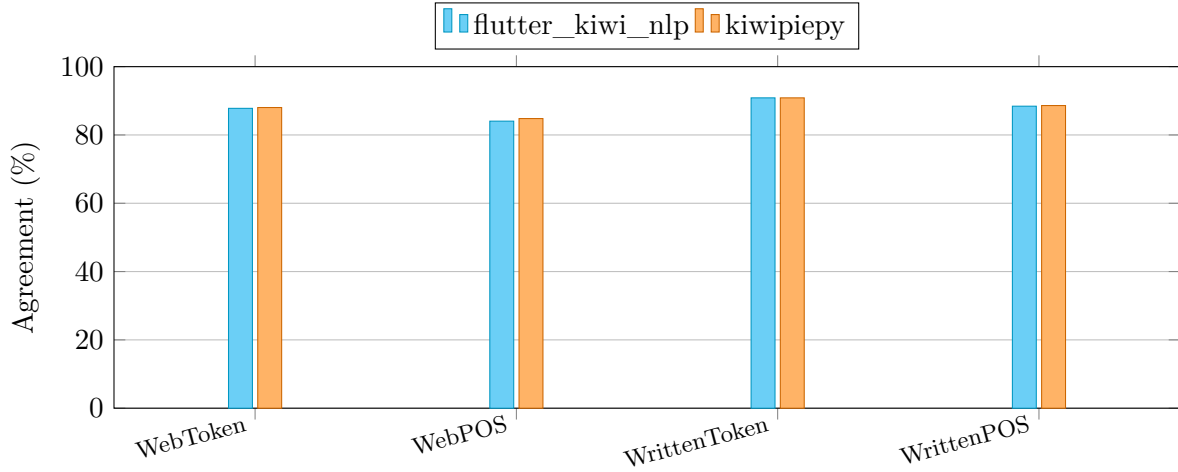


Figure 10: Per-dataset token/POS agreement profile.



Figure 11: Exact-match comparison on gold corpora (strict full-sequence criteria).

14.19 Interpretation of Gold-Corpus Results

Aggregate token and POS agreement are close between runtimes (sub-1pp gap), but strict sentence exact match remains low on both sides. This is expected for Korean morphological pipelines where small boundary/tag normalization differences across wrappers can fail whole-sentence exactness despite high token-level agreement. The combined predicted token totals (Flutter 8,076; Kiwi 8,103) confirm minor segmentation divergence under aligned option bitmasks.

14.20 Recommended Protocol for Fairer Future Comparisons

1. Force same model assets by passing identical absolute `-model-path`.
2. Keep explicit bitmask parity and additionally validate per-sentence output equivalence for sampled cases.
3. Run repeated trials (for example 10+), report mean/median/stddev.
4. Separate measurements into cold init, warm init, and steady-state analysis throughput.
5. Record exact binary/model versions in output payload metadata.

15 Security and Supply-Chain Considerations

15.1 Archive Integrity

Both native and web fallback flows support optional SHA-256 verification, which is critical when model archives are downloaded at runtime.

15.2 Web CSP and CDN Trust Boundary

By default, web runtime bootstrap uses CDN-hosted `kiwi-nlp` module and WASM URLs (`jsDelivr`) unless overridden through `FLUTTER_KIWI_NLP_WEB_MODULE_URL` and `FLUTTER_KIWI_NLP_WEB_WASM_URL`. Production deployments therefore need explicit Content Security Policy (CSP) allowances

for the selected module and WASM origins (or must self-host these artifacts under an already-allowed origin). This CDN trust boundary is distinct from model-archive integrity checks.

15.3 Dynamic Loading Risk

Environment-based dynamic library overrides are useful for controlled deployment, but should be restricted in hardened environments to avoid untrusted path injection.

15.4 Failure Transparency

The plugin intentionally prefers explicit hard failure with context-rich errors instead of silent fallback. This behavior improves diagnosability and reduces risk of undetected incorrect execution.

16 Maintainability Notes

16.1 Separation of Concerns

- Public API and models are compact and typed.
- Runtime-specific logic is isolated in dedicated files (native, web, stub).
- Shared model-file metadata and validation helpers are centralized in `kiwi_model_assets.dart`.

16.2 Testability Hooks

Native analyzer exposes explicit debug hooks for tests:

- binding factory override,
- archive URL/checksum override,
- HTTP client factory override.

These hooks allow deterministic tests for initialization and fallback branches.

17 Testing and Coverage Quality

17.1 Test Suite Scope

As of February 18, 2026, the test suite is organized in layered form:

- Core package layer (`test/`): 63 unit tests in 14 groups across 9 files.
- Example widget/golden layer (`example/test/`): 5 tests, including 3 screenshot(golden) assertions.
- Example integration/acceptance layer (`example/integration_test/`): 3 end-to-end tests.

This yields 71 test declarations across repository test directories. Coverage focus in the core package layer remains strongest on API contract behavior, typed model parsing, option flags, error semantics, and native fallback logic. For macOS desktop execution, the three integration tests pass under serial invocation with explicit device pinning (`-d macos`).

Key tested areas include:

- public API export and unsupported-platform behavior,
- native analyzer lifecycle (`create/analyze/ addUserWord/close`),
- model path resolution and asset/archive fallback branches,
- JSON model decoding for `KiwiToken`, `KiwiCandidate`, and `KiwiAnalyzeResult`,
- option constant composition and exception formatting,
- acceptance flow over analyzer demo UI actions (`analyze/clear/settings/POS-sheet`),
- runtime smoke checks for `create` \rightarrow `analyze` \rightarrow `close`,
- native runtime-path probing on macOS FFI candidate resolution,
- screenshot(golden) stability checks for settings and POS dictionary sheets under a fixed mobile viewport.

17.2 Coverage Snapshot

Using `flutter test -coverage test` on February 18, 2026 (core package layer only):

- Raw line coverage: $384/384 = 100.00\%$.
- Filtered line coverage: $81/81 = 100.00\%$.

Compared with the previous paper snapshot (raw 93.58%), raw line coverage improves by +6.42 percentage points in this revision.



Figure 12: Coverage snapshot used in this paper.

17.3 Why Raw and Filtered Coverage Can Differ

The project includes a filtered coverage gate (`tool/check_coverage.sh`) that excludes:

- generated binding file:
`lib/flutter_kiwi_ffi_bindings_generated.dart`,
- native runtime implementation:
`lib/src/kiwi_analyzer_native.dart`,

- web runtime implementation:
`lib/src/kiwi_analyzer_web.dart`.

The rationale is that parts of these files require runtime/platform integration contexts that are harder to exercise in hermetic unit tests. This makes filtered coverage useful as a strict gate for stable pure-Dart layers, but it must not be interpreted as full end-to-end runtime coverage. In the current snapshot, raw and filtered values are equal (both 100.00%), but the filtered gate remains relevant as a policy mechanism for future regressions.

17.4 Honest Quality Assessment

The present quality state is strong for deterministic unit-level contract tests and includes expanded acceptance/golden coverage in the example app. Remaining risk still concentrates in integration boundaries:

- native dynamic loading and ABI interaction across target platforms,
- web module import/WASM loading under browser/network constraints,
- platform build-hook behavior in real CI/device matrices,
- physical mobile device variability beyond simulator/emulator environments.

Therefore, line coverage should be treated as one quality indicator, not a complete reliability proof. Integration tests on actual target runtimes are the next quality multiplier for this plugin class.

18 Limitations and Future Work

18.1 Current Limitations

- Web runtime depends on module/WASM artifact availability unless module, WASM, and model assets are self-hosted under application-controlled origins.
- Web deployments must satisfy CSP rules for module/WASM fetch origins. Because the default bootstrap uses CDN URLs, this creates an explicit supply-chain boundary unless deployments self-host artifacts.
- Native archive fallback requires network access and writable cache location when model assets are not bundled locally.
- Warm-path throughput can exceed Python-native execution in this revision, but short-session effective throughput still trails because initialization dominates small- N workloads.
- Bridge-level JSON serialization overhead in the native C layer remains a major cost center (dynamic allocation and string copy on hot paths).
- Native execution is in-process via FFI; a crash in dependent native code is fatal to the host Flutter process (no process isolation boundary).
- Platforms outside declared plugin matrix are unsupported.
- Current mobile benchmark rows use simulator/emulator environments (iOS debug simulator, Android emulator), not physical iOS/Android devices.
- Web throughput benchmark rows are not yet included in this report.

18.2 Clarifications on Common Reviewer Questions

- **Could zero-copy or binary bridge further improve throughput and close effective-session gaps?** It is a promising direction and is explicitly planned. Current boundary measurements indicate JSON-path overhead around 13.89–47.93% on warm paths; this suggests meaningful gains, but also implies that non-JSON residual costs may still remain after migration.
- **Does Android init include asset extraction from APK?** Yes, the reported init metric wraps `KiwiAnalyzer.create(...)` and can include model-path resolution and asset-copy work on cache-miss paths, in addition to native bridge/Kiwi initialization.
- **How is Web TTI impacted by WASM/model download size?** First-use web latency is affected by module/WASM/model fetch and runtime instantiation; this report currently does not provide controlled TTI numbers, and therefore treats them as explicit future benchmarking scope.

18.3 Recommended Future Work

1. **Bridge performance redesign:** prototype and A/B evaluate binary payload designs (for example FlatBuffers/Protobuf-style schemas) and typed native transfer paths against current JSON baseline. Report per-phase gains (serialize/parse/call boundary) and safety complexity (ownership/lifetime). To make this roadmap auditable, the next revision will use explicit target thresholds against the current baseline:
 - **Primary target:** at least 40% relative reduction in boundary loss (Table 16) from the current 36.85%/47.93%/19.95% (desktop/iOS/Android) to $\leq 22.11\%/\leq 28.76\%/\leq 11.97\%$.
 - **Primary target:** at least 35% relative reduction in boundary overhead per analysis from 0.0485/0.0891/0.0810 ms to $\leq 0.0315/\leq 0.0579/\leq 0.0527$ ms.
 - **Stretch target:** session-length effective throughput ratio at $N = 1000$ (Table 24) to reach $\geq 0.85x/\geq 0.80x/\geq 0.45x$ (desktop/iOS/Android) from current 0.73x/0.68x/0.33x.
 - **Quality guardrails:** token/POS agreement regression ≤ 0.2 percentage points, no increase in crash/timeout incidence, and peak RSS increase $\leq 10\%$.
 - **Statistical acceptance:** $n \geq 10$ repeated trials per target and 95% CI reporting; an optimization is accepted only when both the mean improvement and the CI lower bound exceed zero.
2. **Benchmark external validity:** add release/profile measurements on physical iOS and Android devices, with repeated runs under thermal-state controls and explicit startup-path annotations (asset-cache hit/miss). The minimum publication matrix is one recent iPhone (A-series) and one recent Android flagship (Snapdragon class), each with $n \geq 10$ repeated trials in release mode and reported p50/p95 plus 95% CI. Include sustained Web throughput and first-use TTI rows under controlled network/cache conditions (cold-cache and warm-cache, plus explicit throttled/unthrottled profiles).
3. **Operational hardening:** provide first-class self-hosting recipes for web module/WASM/model assets, with CSP and integrity examples.
4. **Reliability engineering:** evaluate optional crash-containment architectures for native execution (for example, process separation modes for fault-intolerant deployments).

5. **Compatibility governance:** expand CI validation across Kiwi versions/model families and platform-specific binary packaging combinations.
6. **Quality expansion:** add deterministic regression fixtures for hard tokenization/POS edge cases and broaden integration tests on real target runtimes.
7. **API evolution:** assess optional batch-analysis APIs for high-throughput service scenarios without breaking existing contracts.

19 Conclusion

`flutter_kiwi_nlp` contributes a native-first, cross-platform integration architecture that exposes one stable Dart API while spanning two runtime stacks (native FFI and web WASM). The implementation focus is practical deployment quality: explicit fallback order, reproducible build hooks, typed result contracts, and diagnosable failure behavior.

Empirical evidence in this paper indicates three core outcomes. First, runtime semantics are operationally aligned across targets under a shared API model. Second, the package is suitable for on-device and offline-native inference workflows once required model assets are provisioned locally. Third, current performance trade-offs are transparent: warm-path Flutter throughput can exceed the Python baseline in measured settings (desktop, iOS simulator, Android emulator, and Linux/Windows batch CI rows), while cold-start and short-session effective throughput remain lower. Linux/Windows single-mode rows are still below Python throughput, reinforcing execution-mode dependence as a key systems factor. Boundary-decomposed profiling still identifies a persistent JSON-path penalty (roughly 13.89–47.93% in this report) as a primary optimization target.

This paper intentionally does not claim a new morphology model or SOTA language accuracy contribution. Its contribution is systems engineering: making Kiwi usable in production Flutter applications across Android, iOS, macOS, Linux, Windows, and Web with explicit operational contracts.

For practitioners, the key value is reduced integration risk when one codebase must ship across platforms while preserving deterministic Korean NLP behavior. For researchers and reviewers, the value is a reproducible, implementation-level specification that surfaces both strengths and unresolved constraints.

References

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv preprint arXiv:1810.04805, 2018. [Online]. Available: <https://arxiv.org/abs/1810.04805>. [Accessed: 2026-02-17].
- [2] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” arXiv preprint arXiv:1909.11942, 2019. [Online]. Available: <https://arxiv.org/abs/1909.11942>. [Accessed: 2026-02-17].
- [3] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators,” arXiv preprint arXiv:2003.10555, 2020. [Online]. Available: <https://arxiv.org/abs/2003.10555>. [Accessed: 2026-02-17].

- [4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019. [Online]. Available: <https://arxiv.org/abs/1910.01108>. [Accessed: 2026-02-17].
- [5] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, “TinyBERT: Distilling BERT for Natural Language Understanding,” arXiv preprint arXiv:1909.10351, 2019. [Online]. Available: <https://arxiv.org/abs/1909.10351>. [Accessed: 2026-02-17].
- [6] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers,” arXiv preprint arXiv:2002.10957, 2020. [Online]. Available: <https://arxiv.org/abs/2002.10957>. [Accessed: 2026-02-17].
- [7] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices,” arXiv preprint arXiv:2004.02984, 2020. [Online]. Available: <https://arxiv.org/abs/2004.02984>. [Accessed: 2026-02-17].
- [8] F. N. Iandola, A. E. Shaw, R. Krishna, and K. W. Keutzer, “SqueezeBERT: What can computer vision teach NLP about efficient neural networks?” arXiv preprint arXiv:2006.11316, 2020. [Online]. Available: <https://arxiv.org/abs/2006.11316>. [Accessed: 2026-02-17].
- [9] T. Tambe et al., “EdgeBERT: Sentence-Level Energy Optimizations for Latency-Aware Multi-Task NLP Inference,” arXiv preprint arXiv:2011.14203, 2020. [Online]. Available: <https://arxiv.org/abs/2011.14203>. [Accessed: 2026-02-17].
- [10] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin, “DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference,” arXiv preprint arXiv:2004.12993, 2020. [Online]. Available: <https://arxiv.org/abs/2004.12993>. [Accessed: 2026-02-17].
- [11] F. Fusco, D. Pascual, P. Staar, and D. Antognini, “pNLP-Mixer: an Efficient all-MLP Architecture for Language,” arXiv preprint arXiv:2202.04350, 2022. [Online]. Available: <https://arxiv.org/abs/2202.04350>. [Accessed: 2026-02-17].
- [12] N. Goyal, “A comprehensive study of on-device NLP applications – VQA, automated Form filling, Smart Replies for Linguistic Codeswitching,” arXiv preprint arXiv:2409.19010, 2024. [Online]. Available: <https://arxiv.org/abs/2409.19010>. [Accessed: 2026-02-17].
- [13] S. Wang, A. Shenoy, P. Chuang, and J. Nguyen, “Now It Sounds Like You: Learning Personalized Vocabulary On Device,” arXiv preprint arXiv:2305.03584, 2023. [Online]. Available: <https://arxiv.org/abs/2305.03584>. [Accessed: 2026-02-17].
- [14] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, “KR-BERT: A Small-Scale Korean-Specific Language Model,” arXiv preprint arXiv:2008.03979, 2020. [Online]. Available: <https://arxiv.org/abs/2008.03979>. [Accessed: 2026-02-17].
- [15] S. Park et al., “KLUE: Korean Language Understanding Evaluation,” arXiv preprint arXiv:2105.09680, 2021. [Online]. Available: <https://arxiv.org/abs/2105.09680>. [Accessed: 2026-02-17].
- [16] K. Yang, Y. Jang, T. Lee, J. Seong, H. Lee, H. Jang, and H. Lim, “KoBigBird-large: Transformation of Transformer for Korean Language Understanding,” arXiv preprint arXiv:2309.10339, 2023. [Online]. Available: <https://arxiv.org/abs/2309.10339>. [Accessed: 2026-02-17].

- [17] A. Matteson, C. Lee, Y.-B. Kim, and H. Lim, “Rich Character-Level Information for Korean Morphological Analysis and Part-of-Speech Tagging,” arXiv preprint arXiv:1806.10771, 2018. [Online]. Available: <https://arxiv.org/abs/1806.10771>. [Accessed: 2026-02-17].
- [18] T. Kudo, “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” [Online]. Available: <https://taku910.github.io/mecab/>. [Accessed: 2026-02-17].
- [19] Eunjeon Project, “Eunjeon Korean NLP project page (MeCab-ko lineage),” Blog. [Online]. Available: <http://eunjeon.blogspot.com/>. [Accessed: 2026-02-17].
- [20] Kakao Corp., “Khaiii repository,” GitHub. [Online]. Available: <https://github.com/kakao/khaiii>. [Accessed: 2026-02-17].
- [21] KoNLPy Contributors, “KoNLPy documentation,” [Online]. Available: <https://konlpy.org/en/latest/>. [Accessed: 2026-02-17].
- [22] GitHub Docs, “GitHub REST API documentation.” [Online]. Available: <https://docs.github.com/en/rest>. [Accessed: 2026-02-17].
- [23] bab2min, “Kiwi repository,” GitHub. [Online]. Available: <https://github.com/bab2min/Kiwi>. [Accessed: 2026-02-17].
- [24] Open Korean Text Contributors, “Open Korean Text repository,” GitHub. [Online]. Available: <https://github.com/open-korean-text/open-korean-text>. [Accessed: 2026-02-17].
- [25] SHINEWARE, “KOMORAN repository,” GitHub. [Online]. Available: <https://github.com/shineware/KOMORAN>. [Accessed: 2026-02-17].
- [26] KoNLPy Contributors, “KoNLPy repository,” GitHub. [Online]. Available: <https://github.com/konlpy/konlpy>. [Accessed: 2026-02-17].
- [27] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” arXiv preprint arXiv:1804.07461, 2018. [Online]. Available: <https://arxiv.org/abs/1804.07461>. [Accessed: 2026-02-17].
- [28] A. Wang et al., “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems,” arXiv preprint arXiv:1905.00537, 2019. [Online]. Available: <https://arxiv.org/abs/1905.00537>. [Accessed: 2026-02-17].
- [29] Y. Tay et al., “Long Range Arena: A Benchmark for Efficient Transformers,” arXiv preprint arXiv:2011.04006, 2020. [Online]. Available: <https://arxiv.org/abs/2011.04006>. [Accessed: 2026-02-17].
- [30] C. Coleman et al., “Analysis of DAWNBench, a Time-to-Accuracy Machine Learning Performance Benchmark,” arXiv preprint arXiv:1806.01427, 2018. [Online]. Available: <https://arxiv.org/abs/1806.01427>. [Accessed: 2026-02-17].
- [31] V. J. Reddi et al., “MLPerf Inference Benchmark,” arXiv preprint arXiv:1911.02549, 2019. [Online]. Available: <https://arxiv.org/abs/1911.02549>. [Accessed: 2026-02-17].
- [32] S. S. Chawathe et al., “Tiny Machine Learning: Progress and Futures,” arXiv preprint arXiv:2403.19076, 2024. [Online]. Available: <https://arxiv.org/abs/2403.19076>. [Accessed: 2026-02-17].

- [33] H. B. McMahan et al., “Communication-Efficient Learning of Deep Networks from Decentralized Data,” arXiv preprint arXiv:1602.05629, 2016. [Online]. Available: <https://arxiv.org/abs/1602.05629>. [Accessed: 2026-02-17].
- [34] E. Park and S. Park, “Kiwi: A Study on Developing a Korean Morphological Analyzer,” *Korean Journal of Digital Humanities*, vol. 1, no. 1, pp. 109–135, 2018. [Online]. Available: <https://accesson.kr/kjdh/v.1/1/109/43508>. [Accessed: 2026-02-17].
- [35] bab2min, “Kiwi model/type definitions (include/kiwi/Types.h),” GitHub. [Online]. Available: <https://github.com/bab2min/Kiwi/blob/main/include/kiwi/Types.h>. [Accessed: 2026-02-17].
- [36] bab2min, “Kiwi C++ API and model includes (include/kiwi/Kiwi.h),” GitHub. [Online]. Available: <https://github.com/bab2min/Kiwi/blob/main/include/kiwi/Kiwi.h>. [Accessed: 2026-02-17].
- [37] bab2min, “Kiwi path-scoring implementation (src/PathEvaluator.hpp),” GitHub. [Online]. Available: <https://github.com/bab2min/Kiwi/blob/main/src/PathEvaluator.hpp>. [Accessed: 2026-02-17].

A Appendix A: End-to-End Usage Example

```
import 'package:flutter_kiwi_nlp/flutter_kiwi_nlp.dart';

Future<void> runSample() async {
  final KiwiAnalyzer analyzer = await KiwiAnalyzer.create(
    numThreads: -1,
    buildOptions: KiwiBuildOption.defaultOption,
    matchOptions: KiwiMatchOption.allWithNormalizing,
  );

  final KiwiAnalyzeResult result = await analyzer.analyze(
    'sample sentence for analysis',
    options: const KiwiAnalyzeOptions(topN: 1),
  );

  for (final KiwiCandidate candidate in result.candidates) {
    for (final KiwiToken token in candidate.tokens) {
      // Use token.form, token.tag, token.start, token.length, ...
    }
  }

  await analyzer.addUserWord('newword', tag: 'NNP', score: 1.0);
  await analyzer.close();
}
```

B Appendix B: Native and Web Runtime Equivalence Notes

- Method signatures are intentionally identical across native and web implementations.
- Result model shape is normalized to `KiwiAnalyzeResult` -> `KiwiCandidate` -> `KiwiToken`.

- Lifecycle rules are consistent: use-after-close triggers `KiwiException`.
- Web create-time handling of `numThreads/matchOptions` differs (accepted for parity, not applied at creation).

C Appendix C: Reproducibility Statement

This report describes repository state and behavior as observed on February 18, 2026. Build scripts, dependency versions, defaults, and benchmark numbers may change in future releases. Formalized benchmark/testing pseudocode is provided in Appendix D.

Minimal command sequence used for benchmark reproduction: The measured mobile rows in this revision are simulator/emulator runs; physical device command templates are included below for direct extension.

```
# Desktop baseline (macOS)
uv run python tool/benchmark/run_compare.py \
  --device macos --mode release --trials 3 \
  --warmup-runs 3 --measure-runs 15

# Linux desktop CI row (single mode)
uv run python tool/benchmark/run_compare.py \
  --device linux --mode release --trials 3 \
  --warmup-runs 3 --measure-runs 15 \
  --flutter-analyze-impl token_count \
  --flutter-execution-mode single \
  --output-dir benchmark/results/gha_linux_single_t3_w3_m15_s10

# Linux desktop CI row (batch mode)
uv run python tool/benchmark/run_compare.py \
  --device linux --mode release --trials 3 \
  --warmup-runs 3 --measure-runs 15 \
  --flutter-analyze-impl token_count \
  --flutter-execution-mode batch \
  --output-dir benchmark/results/gha_linux_batch_t3_w3_m15_s10

# Windows desktop CI row (single mode)
uv run python tool/benchmark/run_compare.py \
  --device windows --mode release --trials 3 \
  --warmup-runs 3 --measure-runs 15 \
  --flutter-analyze-impl token_count \
  --flutter-execution-mode single \
  --output-dir benchmark/results/gha_windows_single_t3_w3_m15_s10

# Windows desktop CI row (batch mode)
uv run python tool/benchmark/run_compare.py \
  --device windows --mode release --trials 3 \
  --warmup-runs 3 --measure-runs 15 \
  --flutter-analyze-impl token_count \
  --flutter-execution-mode batch \
  --output-dir benchmark/results/gha_windows_batch_t3_w3_m15_s10

# iOS simulator row (debug in this setup)
uv run python tool/benchmark/run_compare.py \
```

```

--device "iPhone 17" --mode debug --trials 3 \
--warmup-runs 3 --measure-runs 15

# Android emulator row (release)
uv run python tool/benchmark/run_compare.py \
--device <android_emulator_id> --mode release --trials 3 \
--warmup-runs 3 --measure-runs 15

# iOS physical device row (release; requires attached iPhone)
uv run python tool/benchmark/run_compare.py \
--device <ios_physical_device_id> --mode release --trials 3 \
--warmup-runs 3 --measure-runs 15

# Android physical device row (release; requires attached phone)
uv run python tool/benchmark/run_compare.py \
--device <android_physical_device_id> --mode release --trials 3 \
--warmup-runs 3 --measure-runs 15

# Core package unit tests (plugin root)
flutter test test

# Example screenshot(golden) tests (example app)
cd example
flutter test test/kiwi_ui_golden_test.dart

# Example integration/acceptance tests (run serially to avoid build.db lock)
flutter test integration_test/kiwi_runtime_smoke_test.dart -d macos
flutter test integration_test/kiwi_acceptance_flow_test.dart -d macos
flutter test integration_test/kiwi_native_runtime_path_test.dart -d macos
cd ..

# Gold-corpus agreement
uv run python tool/benchmark/gold_corpus_compare.py \
--device macos --mode release

# Wrapper activity quantification snapshot
uv run python tool/benchmark/collect_wrapper_activity.py \
--as-of-date 2026-02-17

# Environment capture used for Table "iOS/Android benchmark environment details"
flutter devices
sw_vers
sysctl -n machdep.cpu.brand_string
sysctl -n hw.ncpu
sysctl -n hw.memsize
xcodebuild -version
xcrun simctl list devices --json
xcrun simctl list runtimes --json
xcrun simctl list devicetypes --json
xcrun simctl spawn <ios_simulator_udid> /usr/sbin/sysctl -n hw.ncpu
xcrun simctl spawn <ios_simulator_udid> /usr/sbin/sysctl -n hw.memsize
adb -s <android_emulator_id> emu avd name
adb -s <android_emulator_id> shell getprop ro.build.version.sdk
adb -s <android_emulator_id> shell cat /proc/meminfo

```

```
adb -s <android_emulator_id> shell cat /proc/cpuinfo  
cat ~/.android/avd/Pixel_9.avd/config.ini
```

Primary generated artifacts:

- benchmark/results/macos_release_t3_json_batch_v2/comparison.md
- benchmark/results/gha_linux_single_t3_w3_m15_s10/comparison.md
- benchmark/results/gha_linux_batch_t3_w3_m15_s10/comparison.md
- benchmark/results/gha_windows_single_t3_w3_m15_s10/comparison.md
- benchmark/results/gha_windows_batch_t3_w3_m15_s10/comparison.md
- benchmark/results/ios_debug_t3_token_count_s10_isolate_worker_v1/comparison.md
- benchmark/results/android_release_t3_token_count_s10_isolate_worker_v1/comparison.md
- benchmark/results/macos_release_t3_json_batch_v2/flutter_kiwi_benchmark_trials.json
- benchmark/results/macos_release_t3_json_batch_v2/kiwipiepy_benchmark_trials.json
- benchmark/results/gold_eval/comparison.md
- benchmark/results/gold_eval/flutter_overall.json
- benchmark/results/gold_eval/kiwipiepy_overall.json
- benchmark/results/wrapper_activity/wrapper_activity.json
- benchmark/results/wrapper_activity/wrapper_activity.md
- per-trial JSON files for each runtime in benchmark/results/.
- screenshot(golden) baselines in example/test/goldens/.

D Appendix D: Benchmark and Test Execution Pseudocode

Figure 13: Cross-runtime benchmark orchestration (`tool/benchmark/run_compare.py`)

Input: device d , mode m , corpus p , trials n , warmup w , measure r

Output: comparison report and per-trial JSON artifacts for Flutter and Python

- 1: Resolve output directory key from (d, m, n, w, r) and options.
- 2: **for** $i \leftarrow 1$ **to** n **do**
- 3: Launch Flutter benchmark target `example/lib/benchmark_main.dart` on (d, m) .
- 4: Wait for marker `KIWI_BENCHMARK_JSON=` in stdout.
- 5: **if** marker is missing on Android **then**
- 6: Scan `adb logcat` and reconstruct payload from base64 chunks.
- 7: **end if**
- 8: Persist Flutter trial payload JSON.
- 9: Run Python benchmark (`tool/benchmark/kiwipiepy_benchmark.py`) with the same corpus and option bitmasks.
- 10: Persist Python trial payload JSON.
- 11: **end for**
- 12: Aggregate trial statistics (mean/stddev, ratio, per-trial snapshots).
- 13: Generate `comparison.md` via `tool/benchmark/compare_results.py`.

Figure 14: Gold-corpus agreement evaluation (`tool/benchmark/gold_corpus_compare.py`)

Input: asset list G , device d , mode m , top- N , build/match options

Output: token/POS agreement metrics and dataset/overall JSON outputs

- 1: **for all** $g \in G$ **do**
- 2: Load tab-separated (`sentence, gold_tokens`) entries from g .
- 3: Run Flutter evaluator target `example/lib/gold_eval_main.dart`.
- 4: Parse `KIWI_GOLD_EVAL_JSON=` payload.
- 5: Run Python evaluator (`kiwipiepy`) with aligned options.
- 6: Align outputs per sentence (top-1 candidate) and compute: token agreement, POS agreement, token-pair exact, sentence exact.
- 7: Persist per-dataset JSON summaries.
- 8: **end for**
- 9: Aggregate micro/macro overall metrics across datasets.
- 10: Emit `benchmark/results/gold_eval/comparison.md`, `flutter_overall.json`, and `kiwipiepy_overall.json`.

Figure 15: Unit-test and coverage gate pipeline (`tool/check_coverage.sh`)

Input: repository root R , coverage threshold τ (default 100%)

Output: pass/fail decision with filtered coverage report

- 1: Execute `flutter test -coverage test`.
- 2: **if** `coverage/lcov.info` does not exist **then**
- 3: **fail** with missing-report error.
- 4: **end if**
- 5: Filter LCOV records, excluding generated/native/web runtime files.
- 6: Compute filtered coverage: `coverage = 100 × LH/LF`.
- 7: Write filtered report to `coverage/lcov.filtered.info`.
- 8: **if** `coverage < τ` **then**
- 9: **fail** coverage gate.
- 10: **else**
- 11: **pass** and report filtered percentage.
- 12: **end if**