EFFOMA TECHNICAL REPORT · EFFOMA-TR-2026-01 · VERSION 2.0

Multilingual by Requirement, Untested by Language: The Evidence Gap in Government AI Procurement

A Nguyễn · Kody Anh Nguyễn · Founder, EFFOMA

3 August 2026

Abstract

Multilingual capability appears in all three Japanese municipal AI tender packages examined: Akashi, Nakano, and Imari. None of the published files reviewed states a pass threshold by language. This report separates language availability—a feature—from language-specific failure bounds—evidence. The tenders require official-source answers, abstention or contact guidance, and operating controls, but their published evaluation plans do not show how the weakest supported language must perform. Akashi's plan for approximately 50 FAQ tests illustrates the statistical gap: even if 50 independent, representative, untouched cases all passed, the exact one-sided 95% upper failure bound would remain 5.82%. It takes 299 zero-failure tests to push that pooled bound below 1%. The report proposes a five-gate launch rule covering evidence, procedure, language, operations, and accountability. It is an assurance proposal, not a validated standard; procurement requirements are not delivered outcomes.

Keywords: multilingual government AI; public-service AI; procurement; reliability; evaluation; language testing; digital government; Japan

The finding

Language availability is a feature. Language-specific failure bounds are evidence. The tenders publish the former, not the latter.

All three municipal packages list multilingual capability. They also require official-source grounding, a safe response when evidence is missing, and some form of contact or department guidance [3–5]. Yet none of the published files reviewed states a minimum pass rate for each required language.

That omission matters because residents do not use a language list. They ask whether they qualify, which document to submit, and what deadline applies. A system can answer in Vietnamese and still mistranslate an exception. It can cite the right page and give the steps in the wrong order. It can display a contact and route the resident nowhere.

National infrastructure does not close that gap. Japan's GENAI program provides a common staff environment, and the national guideline defines a generative-AI system as more than the model alone [1,2]. The municipal tenders are separate procurements, not stages of one national rollout. They show where the service-level burden lands: on the institution that owns the source, procedure, interface, and response path.

What the tenders publish

Package	Multilingual requirement	Evidence and safe failure	Published evaluation gap
Akashi [3]	Multilingual answers are mandatory; the reviewed feature sheet does not name a fixed language set	Answers are limited to city-designated sources; source title, link, and relevant passage are shown; the service must not guess when evidence is missing or unreliable; unresolved feedback can point to a city contact	About 50 city-prepared FAQs are planned for pre-launch testing, with detected errors corrected. No per-language sample size or pass threshold is stated
Nakano [4]	English, traditional and simplified Chinese, and Korean are mandatory; Tagalog, Vietnamese, Burmese, and easy Japanese are optional	Answers show the source URL, file, or passage; unsupported questions trigger abstention or contact guidance	The questions, method, items, and pass standard are deferred to a post-award plan. No published sample size or per-language threshold is stated
Imari [5]	The feature sheet gives at least eight same-language examples: English, French, German, Chinese, Korean, Thai, Russian, and Vietnamese	Answers use official site or Q&A content; absent or low-confidence information triggers abstention and contact guidance	A test environment and report are required. No per-language pass rate is stated

The packages contain real operating requirements. Akashi calls for monthly quality and usage reporting, restrictions on training from resident questions, prompt-injection controls, domestic data handling, and continuity measures [3]. Nakano specifies a 99.9% availability target, normally five-second responses, five-minute monitoring, and incident notice within one hour [4]. Imari requires logs and weekly source crawling [5]. These are contract or planning requirements, not observed performance.

The pattern is precise: **multilingual access is specified; multilingual assurance is not.** “Supports eight languages” says what the interface accepts. It does not say how often policy meaning survives, whether critical errors block launch, or how uncertainty changes for the smallest language test set.

Research in adjacent settings shows why feature counts are insufficient. A small emergency-translation study found recurring problems in terminology, register, and contextual ambiguity [6]. A real-world text-to-911 case study questioned what a claimed 55-language capability establishes at the point of use [7]. JourneyBench found that policy adherence and task completion are distinct evaluation targets [8]. Interviews about a metropolitan 311 chatbot surfaced concerns about interpretation, transparency, and social context [9]. None of these studies measures the Japanese services; each exposes a failure type that a language list cannot rule out.

The 50-test trap

Akashi plans for approximately 4,000 users and 10,000 question-response interactions per month. Its specification calls for roughly 50 city-prepared FAQs before release and correction of the errors found [3]. It does **not** report 50 clean passes.

Consider the most optimistic hypothetical: all 50 cases are frozen before testing, independent, representative of production use, scored against a predeclared binary failure rule, and all pass. For zero failures in n trials, the exact one-sided 95% upper confidence bound is [10]:

```
p_upper = 1 - 0.05^(1/n)
```

Zero-failure tests	One-sided 95% upper failure bound	What the pooled result can support
50	5.82%	Failure rate below 5.82%, under the stated assumptions
299	0.997%	Failure rate below 1%, under the stated assumptions

So **50/50 cannot establish a failure rate below 1%**. And even 299 pooled tests would not establish that every language is below 1%. A strong Japanese result can hide a weak Vietnamese result; low-risk opening-hours questions can hide failures on eligibility or deadlines.

Real test sets are usually less tidy than the formula assumes. Paraphrases can share the same source and retrieval path. FAQ cases are rarely a random sample of production traffic. Cases used to tune the service are no longer untouched acceptance evidence. The calculation is not a universal sample-size prescription. It is a warning against making a strong reliability claim from a small, pooled demonstration.

A weakest-link launch rule

The missing contract can be stated without inventing one composite score. Five gates must pass:

```
Ready = Evidence ∧ Procedure ∧ Language ∧ Operations ∧ Accountability
```

Gate	Pass question	Proof a buyer should require
Evidence	Is each substantive claim supported by a current official source, or does the system abstain safely?	Citation-support tests, source owners, freshness limits, and unanswerable cases
Procedure	Are the rule, sequence, deadline, and authority boundary correct?	Risk-stratified pathway tests and critical-error rules reviewed by service owners
Language	Does meaning and task success hold in every declared language?	Separate counts, failures, and confidence bounds by language and risk class
Operations	Does the service remain safe through source changes, attacks, outages, and model updates?	Source-update drills, security tests, monitoring, rollback, and continuity exercises
Accountability	Can every material answer, incident, correction, and contact path be traced to an owner?	Versioned logs, named owners, response targets, incident records, and correction history

The gates are an author proposal derived from the observed requirements. They have not been validated as a standard. Their value is the decision rule: one failed mandatory gate blocks the claim. High uptime cannot compensate for wrong eligibility guidance. Excellent Japanese performance cannot compensate for an untested language advertised to residents.

A citation is not evidence unless it is current, supports the answer, and has an owner. Contact guidance is not a completed handoff. A robust service must also show where the request goes, who owns it, when a response is due, and what happens when capacity is exhausted.

What buyers should ask

Before launch, a buyer should be able to inspect five artifacts:

1. **A declared boundary:** supported languages, tasks, risk classes, official sources, exclusions, and named owners.
2. **A frozen acceptance plan:** untouched cases, sampling rationale, scoring rules, critical-failure rules, and target confidence bounds.
3. **Stratified results:** counts and uncertainty for every mandatory language, risk class, device path, abstention path, and contact route.
4. **Exercised failure paths:** source-update, prompt-injection, outage, queue-load, rollback, and manual-fallback drills.
5. **A change record:** versioned models, prompts, sources, tests, incidents, corrections, and revalidation triggers.

Three questions expose weak procurement language quickly:

- “Supports eight languages” according to which per-language cases, failure definitions, sample sizes, and bounds?
- Who owns each cited source, and how quickly must the answer change when that source changes?
- Where does contact guidance lead near closing time, and what happens when that destination is unavailable?

Replacing a model does not update a stale benefits page, resolve a contradictory procedure, or staff a multilingual contact channel. Those are service operations. They are also what residents experience.

Method and limits

This exploratory report reviews the published procurement packages for resident-facing generative-AI services in Akashi, Nakano, and Imari, available by 3 August 2026. Each package was read across its specification and feature or non-functional requirements, not just its landing page. Two national sources supplied policy context [1,2]. Four research publications supplied conceptual comparison [6–9].

Claims were extracted when a file stated a language, source, abstention, contact, test, monitoring, security, or change-control requirement. Negative findings are deliberately narrow: “not stated in the published files reviewed” does not mean that a later plan, contract, or deployed service lacks the control.

The sample is purposive and does not represent Japan's municipal market. Tender requirements are intended controls, not evidence of implementation quality or resident outcomes. One researcher conducted the review without independent bilingual or inter-rater validation. The five-gate rule is unvalidated, and the statistical illustration assumes fixed, independent, representative trials with a predeclared failure definition.

Conclusion

The three tenders make multilingual access visible. They do not publish the evidence needed to judge the weakest language.

That is the procurement gap. A language list belongs in a feature table. A reliability claim requires separate cases, failure rules, sample sizes, and uncertainty for every language and risk class the service promises to support.

Government AI should be evaluated as the service residents use—not the model a vendor supplies. Until the weakest mandatory language, procedure, operating control, and response path pass, the service has not demonstrated readiness.

Declarations

Review status. This independently published technical report has not undergone external peer review.

Author contribution. Kody Anh Nguyễn conceived and directed the report and is responsible for its review, interpretation, and publication.

Competing interests. The author is the founder of EFFOMA, which develops multilingual AI systems in related service settings. No EFFOMA deployment is used as validation evidence here.

Funding. No external funding was received.

Ethics. The report analyzes public documents and published research. It involves no human participants, personal data, or intervention.

Data and materials. All source packages are publicly available at the URLs below. The equation and inputs needed to reproduce the zero-failure bounds appear in this report.

AI assistance. Generative AI tools assisted with source discovery, calculation checking, drafting, and editing under the author's direction. The author remains responsible for verification and publication.

How to cite this report

A Nguyễn (Nguyễn Hưng Tuấn Anh / Kody Anh Nguyễn). (2026). *Multilingual by requirement, untested by language: The evidence gap in government AI procurement* (EFFOMA Technical Report EFFOMA-TR-2026-01, Version 2.0). EFFOMA. https://effoma.com/blog/reliability-framework-multilingual-government-ai-services

```
@techreport{nguyen2026multilingual,
  author      = {Nguyễn, Anh},
  title       = {Multilingual by Requirement, Untested by Language: The Evidence Gap in Government AI
Procurement},
  institution = {EFFOMA},
  number      = {EFFOMA-TR-2026-01},
  year        = {2026},
  month       = {August},
  type        = {Technical Report},
  url         = {https://effoma.com/blog/reliability-framework-multilingual-government-ai-services},
  note        = {Version 2.0; author also published as Nguyễn Hưng Tuấn Anh and Kody Anh Nguyễn}
}
```

References

1. Digital Agency, Government of Japan. Government AI “GENAI”. 2026. https://www.digital.go.jp/en/policies/genai. Accessed 2026 Aug 3.
2. Digital Agency, Government of Japan. Guideline for Japanese Government Procurements and Utilizations of Generative AI, Version 2.0 [PDF]. 2026 Jun 12. https://www.digital.go.jp/assets/contents/node/information/field_ref_resources/decb64eb-f26e-41cb-8d37-f3dd173108b8/47764804/20260612_resources_standard_guidelines_guideline_04.pdf. Accessed 2026 Aug 3.

3. Akashi City. Generative AI Chatbot Procurement Package [in Japanese]. 2026. Procurement page: https://www.city.akashi.lg.jp/soumu/digital/nyusatsu/2026/2026_aichatbot.html. Service specification: https://www.city.akashi.lg.jp/documents/40234/shiyousho_aichatbot.pdf. Functional requirements: https://www.city.akashi.lg.jp/documents/40234/kinouyouken_aichatbot_2.xlsx. Non-functional requirements: https://www.city.akashi.lg.jp/documents/40234/20260723155958.pdf. Accessed 2026 Aug 3.
4. Nakano City. Procurement of a Generative AI Search Engine for the Ward Website [in Japanese]. 2026 Jul 22. Procurement page: https://www.city.tokyo-nakano.lg.jp/jigyosha/nyusatsu/jigyousyasentei-bosyu/seiseiAIkensaku.html. Specification: https://www.city.tokyo-nakano.lg.jp/jigyosha/nyusatsu/jigyousyasentei-bosyu/seiseiAIkensaku.files/90siyosyo.pdf. Functional requirements: https://www.city.tokyo-nakano.lg.jp/jigyosha/nyusatsu/jigyousyasentei-bosyu/seiseiAIkensaku.files/91kinoyoken.xlsx. Accessed 2026 Aug 3.
5. Imari City. Public Proposal for an AI Chatbot System [in Japanese]. 2026; updated 2026 Jul 22. Procurement page: https://www.city.imari.lg.jp/26823.htm. Specification: https://www.city.imari.lg.jp/secure/38663/siyousyo_AICB.pdf. Functional requirements: https://www.city.imari.lg.jp/secure/38663/08_kinouyouken_AICB.xlsx. Accessed 2026 Aug 3.
6. Staiano MC, Han L, Monti J, Chiusaroli F. ITALERT: Assessing the Quality of LLMs and NMT in Translating Italian Emergency Response Text. In: Proceedings of Machine Translation Summit XX. 2025. https://aclanthology.org/2025.mtsummit-1.43/.
7. Court S, Downing L, Elsner M. LLMs in the Real World: Evaluating “AI” in Emergency Contexts. In: Findings of the Association for Computational Linguistics: ACL 2026. 2026. https://aclanthology.org/2026.findings-acl.1779/.
8. Balaji S, Mishra P, Sachdeva A, Agrawal S. Beyond IVR: Benchmarking Customer Support LLM Agents for Business-Adherence. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics: Industry Track. 2026. https://aclanthology.org/2026.eacl-industry.15/.
9. Zhou J, Shen R, You Y, DiSalvo C, Dombrowski L, MacLellan C. Improving Public Service Chatbot Design and Civic Impact: Investigation of Citizens' Perceptions of a Metro City 311 Chatbot. In: Proceedings of the ACM Designing Interactive Systems Conference. 2025. doi:10.1145/3715336.3735831. https://doi.org/10.1145/3715336.3735831.
10. Hanley JA, Lippman-Hand A. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA. 1983;249(13):1743–1745. doi:10.1001/jama.1983.03330370053031. https://doi.org/10.1001/jama.1983.03330370053031.