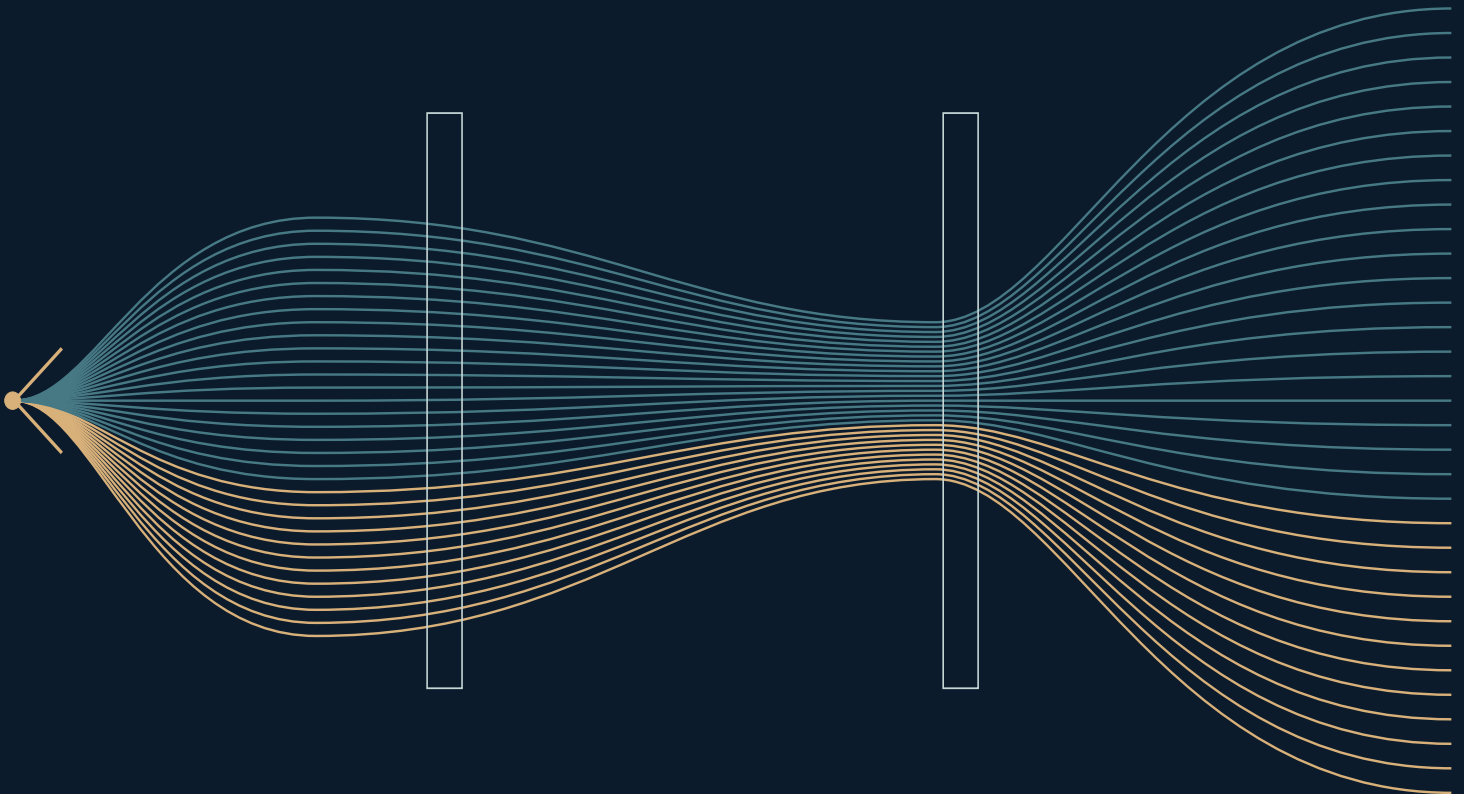




REGION / OCTOBER 2026

Arabic quality is a service promise

Designing AI that works across language, evidence and the complete customer journey



THE DECISION IN BRIEF

Arabic quality is the ability to complete the intended service while preserving meaning, evidence and consequential facts across the supported language varieties and channels.

- 01

Specify the service promise: users, tasks, channels and language varieties. A formal-language benchmark cannot establish whether a customer can complete a mixed-language claim or dispute.
- 02

Compare direct, translated and hybrid routes on the same cases. Judge the preserved facts, accepted outcome and correction effort as well as the fluency of the reply.
- 03

Keep business rules consistent and adapt interpretation and communication where required. One product owner should be accountable for the full Arabic journey, including clarification and human handover.

THE MAYDEN VIEW

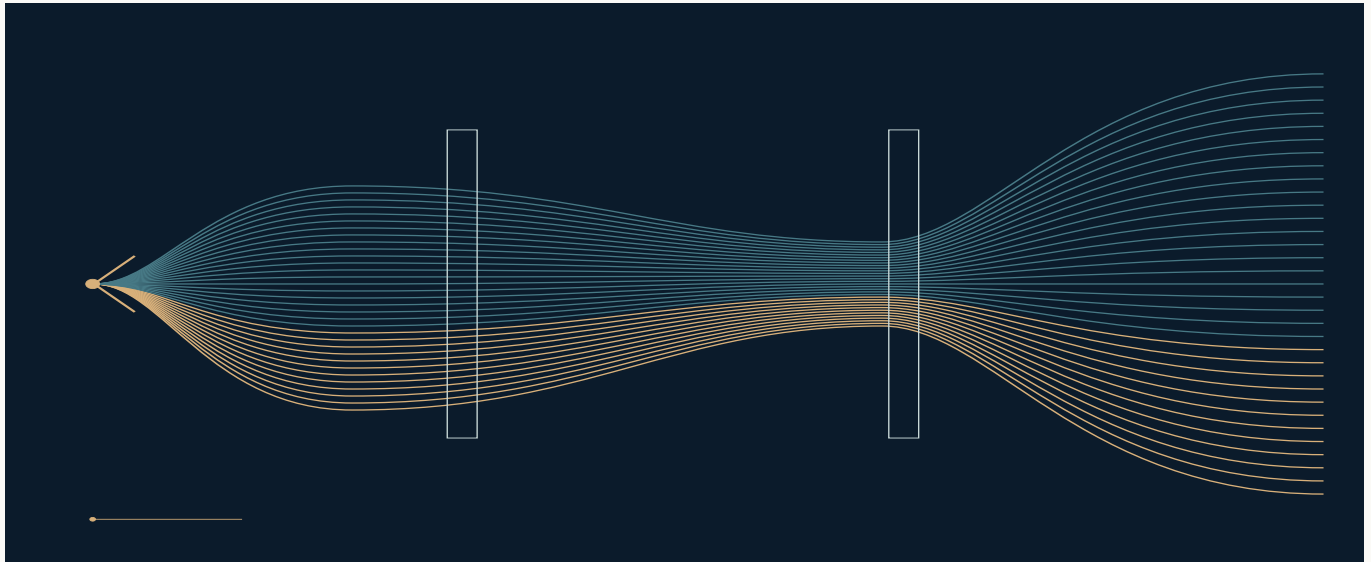
Arabic quality is proved by the completed service, not the fluency of the first reply.

INSIDE THIS PERSPECTIVE

The argument at a glance	3
01 / Specify the Arabic service promise	4-5
Exhibit 01 / Arabic benchmarks cover different parts of the service	6
02 / Compare routes through the same task	7
Exhibit 02 / Test the captured claim, not only the fluent answer	8
03 / Preserve the facts to submission	9-10
Exhibit 03 / Compare complete routes through the same task	11
04 / Invest in the observed service gap	12
The leadership agenda	13
Method and references	14

ONE DECISION, FOUR CONNECTED QUESTIONS

Design for the language of use



RIGHT-TO-LEFT EVIDENCE FLOW / CONCEPTUAL ARTWORK

01 Specify the Arabic service promise
A fluent reply can still produce the wrong record

02 Compare routes through the same task
Separate comprehension from transaction accuracy

03 Preserve the facts to submission
Preserve meaning through the interface

04 Invest in the observed service gap
Share business rules; specialise the language work

Arabic quality is proved by the completed service, not the fluency of the first reply.

CHAPTER 01

01

Specify the Arabic service promise

A fluent reply can still produce the wrong record

A customer describes a claim in colloquial Arabic, names a product in English and attaches a formal document. The service must preserve the event date, identifier and any negation through to submission. A fluent reply can still produce the wrong claim. The investment decision is therefore which language route completes the promised task accurately, at an acceptable operating cost.

ArabicMMLU contains 14,575 multiple-choice questions across 40 tasks in Modern Standard Arabic, sourced from educational examinations across several regions. ^[1] It provides a valuable knowledge-understanding test with a defined

population. It does not establish that a model can resolve an informal complaint, recognize a spoken account number or explain an ambiguous policy. A benchmark should inform a shortlist and an evaluation design; it should not substitute for evidence that a customer can complete the intended task.

Define a testable Arabic service promise: who can complete which task, through which channels and language varieties, with what route to clarification or a person. This makes launch scope explicit and gives product teams a decision they can evaluate. A general claim to support Arabic leaves the consequential choices unresolved.

CHAPTER 01

01

Specify the Arabic service promise

Test the language varieties the service promises

Arabic services operate across formal writing, local speech, mixed-language expressions and differing conventions for names and numbers. AraBench's dialectal Arabic-to-English translation suite includes four coarse, fifteen fine-grained and twenty-five city-level dialect categories across multiple genres. ^[2] Those counts describe the benchmark's organization, not a definitive taxonomy of Arabic speakers. Their value for service design is to make variation visible instead of treating one formal test set as representative of every interaction.

The MADAR shared task likewise distinguishes travel-domain dialect identification from identification using Twitter users, with city- and country-level labels. ^[3] Recognizing a dialect is a narrower task than delivering an appropriate service in it. A company should not infer a user's identity, preferences or eligibility from a model's dialect guess. Where language selection

matters, offer a clear choice or ask a practical clarification instead of making a consequential assumption.

ArabCulture-Dialogue adds a useful dimension to Arabic evaluation. The April 2026 preprint contains 3,471 dialogue cases in paired Modern Standard Arabic and dialect versions, covering 13 countries; its construction involved 26 annotators. The authors acknowledge incomplete country coverage, variation within national dialects and translation effects. ^[8] These are evaluation resources, not proof that a customer service works well in every market.

For a claims service, test the same intention through formal Arabic, the dialects in the launch market and mixed-language input. Keep cultural comprehension separate from the accuracy of the eventual claim record. A model can choose a plausible reply yet still reverse a date, omit an exception or route the case incorrectly. mayden proposes one release decision with separate evidence for comprehension, transaction accuracy and a usable escalation path.

RESEARCH COMPARISON

Arabic benchmarks cover different parts of the service

Study-design comparison

ArabicMMLU, 2024	Published scope 14,575 questions; 40 MSA tasks	What it leaves unresolved Informal conversation and transaction completion
AraBench, 2020	Published scope 25 city-level dialect categories	What it leaves unresolved Full service delivery beyond translation
LAraBench, 2024	Published scope 33 tasks; 61 datasets	What it leaves unresolved An enterprise journey under its own policies
BALSAM, 2025	Published scope 78 tasks; 14 categories	What it leaves unresolved Local users, channels and business consequences
THE MANAGEMENT IMPLICATION Build a local evaluation around the gaps between benchmark tasks and the promised service.		

Counts describe different units and must not be plotted as comparable model performance. Historical suites inform coverage, not current procurement rankings. Sources: [1], [2], [4], [5]

CHAPTER 02

Compare routes through the same task 02

Separate comprehension from transaction accuracy

LAraBench evaluated 33 tasks across 61 public datasets, including language and speech processing, in its 2024 study. ^[4] Its breadth helps identify which parts of the service require separate testing. Transcription, classification, question answering and text generation can have different failure patterns. A service that performs well on one may still fail where its operating journey joins them. Keep those components visible in the evaluation rather than collapsing them into one Arabic score.

For a voice-based support service, measure whether the transcript preserves the information required to act. A transcript can contain several harmless wording differences while preserving the correct request. It can also look fluent while changing a negation, a date or a product code. Review those fields against the source and intended action. The metric should reflect the cost of misunderstanding, with language specialists and service owners jointly defining what counts as a material error.

BALSAM's 2025 publication describes 78 tasks in fourteen categories and a blind evaluation platform, with separate development and test examples. ^[5] Its design addresses breadth and evaluation exposure. Enterprise teams can adopt the principle of holding back fresh cases without claiming that their private set is a representative benchmark. Preserve cases that were not used to tune prompts or choose a model, and refresh them as the service encounters new requests.

Choose the route using matched business cases

There are at least three practical routes through a bilingual service. The system can process Arabic directly, translate through an English intermediate representation, or combine language-specific components with a shared business workflow. Each route changes where meaning can be lost and where specialist review is required. Choose through a task-level comparison. The strongest general-purpose model may be preferable for one service; a specialized component may be better for another. Architecture should follow observed performance and operating constraints.

The original Jais and Jais-chat paper describes Arabic-centric foundation and instruction-tuned models trained with Arabic, English and code. ^[6] It demonstrates a concrete research approach to language-specific capability; its 2023 model comparisons are not current procurement rankings. The useful question is what a candidate model contributes to the service today, under the organization's data, deployment and review conditions. Brand association with a language or region cannot answer that question.

Run the alternatives on matched cases and retain the same acceptance criteria. Count missing qualifiers, altered names, unsupported conclusions, clarifications and specialist corrections. Include the full cost of translation, inference, review and recovery. Record whether any routing rule sends confidential material to a different processing environment. A route that improves language quality may still be unsuitable under a separate operational constraint. Make that trade-off explicit rather than hiding it inside a single technical score.

ILLUSTRATIVE DECISION TOOL

Test the captured claim, not only the fluent answer

English specification for a proposed Arabic claims-intake test

Synthetic test condition	Expected case record	Release consequence if lost
Payment negation	Payment received: no; preserve the source statement	Block submission until corrected
Two dates	Incident: 12 September 2026; invoice date in its own field	Clarify ambiguous calendar or conflicting labels
Claim identifier	CL-1042 exactly; no character substitution	Request confirmation if uncertain
Coverage question	No coverage decision; specialist retains authority	Escalate rather than invent permission
Human handover	Original material, confirmed fields and open questions together	A transfer that makes the customer restart is not completion

THE MANAGEMENT IMPLICATION

Evaluate the same expected fields across direct, translation and hybrid routes before selecting the service design.

All facts are hypothetical. Arabic utterances, dialect coverage and acceptance decisions require competent local review. No native review, field validation or comparative performance is claimed. Sources: [4], [7]

CHAPTER 03

03

Preserve the facts to submission

Preserve meaning through the interface

W3C's Arabic and Persian Layout Requirements describes right-to-left writing, bidirectional text and layout concerns. The October 2025 publication is a Group Draft Note, not an endorsed compliance standard. ^[7] It is a useful technical reference for a product problem that executives should understand: a correct underlying answer can become difficult to use when names, numbers, controls or document references are presented badly.

Test the complete journey on the devices and assistive technologies relevant to the audience. Include mixed-script account identifiers, dates, currency, error messages, attachments and downloaded records. Check whether copying a reference preserves its intended order and

whether a customer can find the next action without changing language. These are proposed acceptance checks requiring direct usability testing. A mirrored navigation bar provides little evidence about the usability of a consequential form or decision notice.

Content ownership matters as much as rendering. Establish an approved vocabulary for products, policy terms and customer actions. Preserve distinctions that carry legal or operational meaning, while allowing natural explanations for users. A literal translation can be technically faithful yet unhelpful; an elegant paraphrase can accidentally remove a condition. Give the bilingual reviewer access to the original policy and the task context, with a clear process for resolving ambiguous terms rather than improvising them repeatedly.

CHAPTER 03

03

Preserve the facts to submission

Keep the claim's facts intact from intake to submission

Consider a fictional insurer designing a claims-intake assistant. Customers may describe an incident in a local dialect, submit an English invoice and receive formal Arabic correspondence. The assistant can collect information and prepare a case summary; it cannot decide coverage. This scope creates a useful comparison between direct Arabic processing and a translation-based route while keeping the consequential decision with an authorized specialist. No performance or client outcome is assumed in this example.

Start with the English operational specification in Exhibit 02. The fictional customer says that no payment was received, gives an incident date of 12 September 2026 and supplies claim identifier CL-1042. An invoice contains a different billing date. The expected record must retain the negation, select the incident date from its labelled source and preserve the identifier exactly. It must not silently substitute the invoice date or infer coverage. These are synthetic facts used to define acceptance, not an Arabic-language test result.

A competent Arabic reviewer should then author and check matched utterances in the language varieties and channels the service intends to support. Keep the underlying facts and expected fields constant while varying the expression.

Record who reviewed each case, which variety it represents and any unresolved interpretation. This report supplies neither those Arabic utterances nor a claim of native review. A date without a clear calendar or an identifier that cannot be read confidently should trigger clarification or human handover, not a guessed normalization.

Suppose one route requires fewer clarifications but produces more specialist corrections to mixed-language documents. That result would not automatically justify selecting either route. Examine whether the corrections concern decisive facts or cosmetic phrasing, whether they are concentrated in identifiable document types and whether routing can isolate those cases. The options include a document-specific path, a narrower launch scope or additional review. The test should reveal which operational choice the evidence supports.

Let customers inspect and correct the captured facts before submission in a language they can use. Preserve the original material alongside the approved summary under the applicable retention and access controls. When transferring to a person, retain context so the customer does not need to repeat the story. This is a proposed service design. Its commercial hypothesis is that accurate intake and effective handover may reduce avoidable rework; the organization must measure that outcome before claiming it.

ILLUSTRATIVE SCENARIO

Compare complete routes through the same task

Architecture options; no ranking implied

Route	Potential advantage	Question to test
Direct Arabic	Fewer translation transformations	Does task-specific evidence use meet the standard?
Translation intermediate	Access to an established shared workflow	Which qualifiers or identifiers change?
Hybrid components	Targeted handling of document or channel needs	Can routing and ownership remain manageable?

THE MANAGEMENT IMPLICATION

Choose using accepted outcomes, review cost and deployment constraints.

mayden analysis. Benefits are hypotheses, not measured outcomes. Sources: [4], [6]

CHAPTER 04

04

Invest in the observed service gap

Share business rules; specialise the language work

Language-specific work has a cost: duplicated operations, fragmented releases and specialist capacity. Arabic-first design need not require a separate foundation model or every dialect at launch. It requires an explicit supported experience and evidence from the people and tasks it is intended to serve.

Separate reusable business rules from language-dependent interpretation and communication. Keep eligibility, transaction limits and access rights consistent where policy is consistent; adapt vocabulary, clarification and interaction to the user. mayden proposes this separation as an option to test, not a universally superior architecture. Its value depends on scale, change frequency and the ability to operate the components.

Prioritize gaps by consequence and observed demand. A rare pronunciation issue affecting a critical identifier may deserve attention before a frequent stylistic preference. Maintain a supported-scope statement and an effective human route for cases outside it. Avoid presenting an untested dialect as covered merely because the model occasionally produces convincing examples. A credible limited promise is a stronger starting point for expansion than an expansive claim that users discover is unreliable.

Invest where the Arabic journey loses meaning

Review Arabic service quality through completion, correction, escalation and unresolved outcomes for the supported segments. Compare matched cases where possible and account for differences in case complexity. A raw performance gap does not establish a language effect; small samples should remain visibly uncertain. Use the review to find the specific journey that needs repair.

Give one product owner the authority and budget to maintain the complete experience. Language and domain specialists need access to source policies, terminology decisions and new failure cases. Reassess when the model, policy or channel changes. Launch approval is the beginning of the service commitment, not its final translation check.

Choose the next investment from the failure pattern. If decisive facts are lost in translation, change or narrow that route. If the record is correct but customers cannot submit it, fix the interface or handover. If specialists repeatedly correct policy terms, repair terminology ownership. mayden's differentiation hypothesis is better completed work and lower avoidable effort; measure those outcomes before claiming a commercial advantage.

FROM INSIGHT TO ACTION

Decisions to take forward

ARABIC SERVICE OWNER AND COMPETENT LANGUAGE REVIEWERS

Approve only the language and task scope actually reviewed and tested.

Evidence to require: Matched cases with reviewed variety, channel and expected fields

ENGINEERING AND CLAIMS SPECIALISTS

Select on decisive-fact preservation and complete effort, not fluency alone.

Evidence to require: Same-case direct, translation and hybrid records, corrections and handovers

OPERATIONS OWNER

Retain human coverage authority and narrow launch where clarification or recovery is unreliable.

Evidence to require: Customer correction and context-preserving transfer exercise

mayden connects senior judgement, research and engineering to turn strategic choices into working capabilities.

EVIDENCE YOU CAN INSPECT

Research notes

Method and boundaries

Narrative comparison of seven opened primary Arabic NLP research records and a W3C draft note, accessed 10 October 2026, with original mayden service-design analysis.

Sources cover native MSA knowledge, dialect translation and identification, language and speech tasks, blind evaluation, model development and script layout.

No new benchmark runs, user interviews or market-size estimates. Historical model results are not current rankings. Benchmark units are incompatible for pooled comparison. Proposed service design requires local validation. The English claims-intake specification is synthetic; no Arabic test utterances were authored or native-reviewed, and no route was experimentally evaluated.

Primary sources and author research

[1] ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic

Koto et al.; Findings of ACL / 2024-07-30, arXiv v2. Abstract. Historical Modern Standard Arabic multiple-choice benchmark drawn from educational exams; 14,575 questions across 40 tasks do not establish informal service or dialect performance.

[Read the source](#)

[2] AraBench: Benchmarking Dialectal Arabic-English Machine Translation

Sajjad et al.; COLING / 2020-12. Abstract. Dialectal Arabic-English translation benchmark with four coarse, 15 fine and 25 city categories; not a complete taxonomy of Arabic users or all conversational tasks.

[Read the source](#)

[3] The MADAR Shared Task on Arabic Fine-Grained Dialect Identification

Bouamor et al.; WANLP / 2019-08. Abstract. Dialect-identification shared task using travel-domain and Twitter material; not a measurement of end-to-end customer-service quality or individual identity.

[Read the source](#)

[4] LArABench: Benchmarking Arabic AI with Large Language Models

Abdelali et al.; EACL / 2024-03. Abstract. Historical language-and-speech benchmark covering 33 tasks and 61 datasets; coverage counts are not current model performance.

[Read the source](#)

[5] BALSAM: A Platform for Benchmarking Arabic Large Language Models

Al-Matham et al.; ArabicNLP / 2025-11. Abstract. Benchmark platform with 78 tasks across 14 categories and 52,000 examples; test coverage does not by itself demonstrate model quality or production suitability.

[Read the source](#)

[6] Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models

Sengupta et al. / 2023-09-29, arXiv v2. Abstract. First-party model-development research from 2023; training and reported evaluation are not an independent current procurement recommendation.

[Read the source](#)

[7] Arabic & Persian Layout Requirements

W3C Internationalization Working Group / 2025-10-02, Group Draft Note. Status; text direction; bidirectional text; layout. A W3C Group Draft Note dated 2 October 2025, not an endorsed W3C standard or a compliance certificate; technical layout guidance is still useful.

[Read the source](#)

[8] Cultural Benchmarking of LLMs in Standard and Dialectal Arabic Dialogues

Al Kautsar et al.; arXiv / 2026-04-30, v1. Sections 3.1-3.4; Table 1; Limitations; Appendix B. Preprint; partial geographic and within-country dialect coverage; translationese; not an end-to-end service benchmark.

[Read the source](#)