How to evaluate AI for multilingual content

Phrase

FORRESTER
WAVE
LEADER 2025

About this guide

This guide is intended to help customers of Phrase understand how to set up evaluations of multilingual content in order to understand best fit for AI solutions in their workflows. It is by no means exhaustive; evaluation of multilingual content is complex and success depends heavily on carefully defining your use case and success criteria.

AI has tremendous benefits, but is always a compromise. No AI solution will meet expectations 100% of the time and in many cases, replacing part or all of a workflow with AI carries some level of risk. This is particularly true in multilingual content. Despite the promise of AI systems in English language tasks, general performance of even the best AI systems is not uniform across all languages and content types.

The goal of any evaluation of AI is to understand to what degree the performance of the systems matches our tolerance for risk. This guide will give you the tools to establish what that tolerance is and how to determine whether AI-based approaches are a good fit for your multilingual content workflows.

Reality check:

AI won't hit targets 100% of the time. This evaluation tells you whether performance meets your tolerance for risk.

The methodology

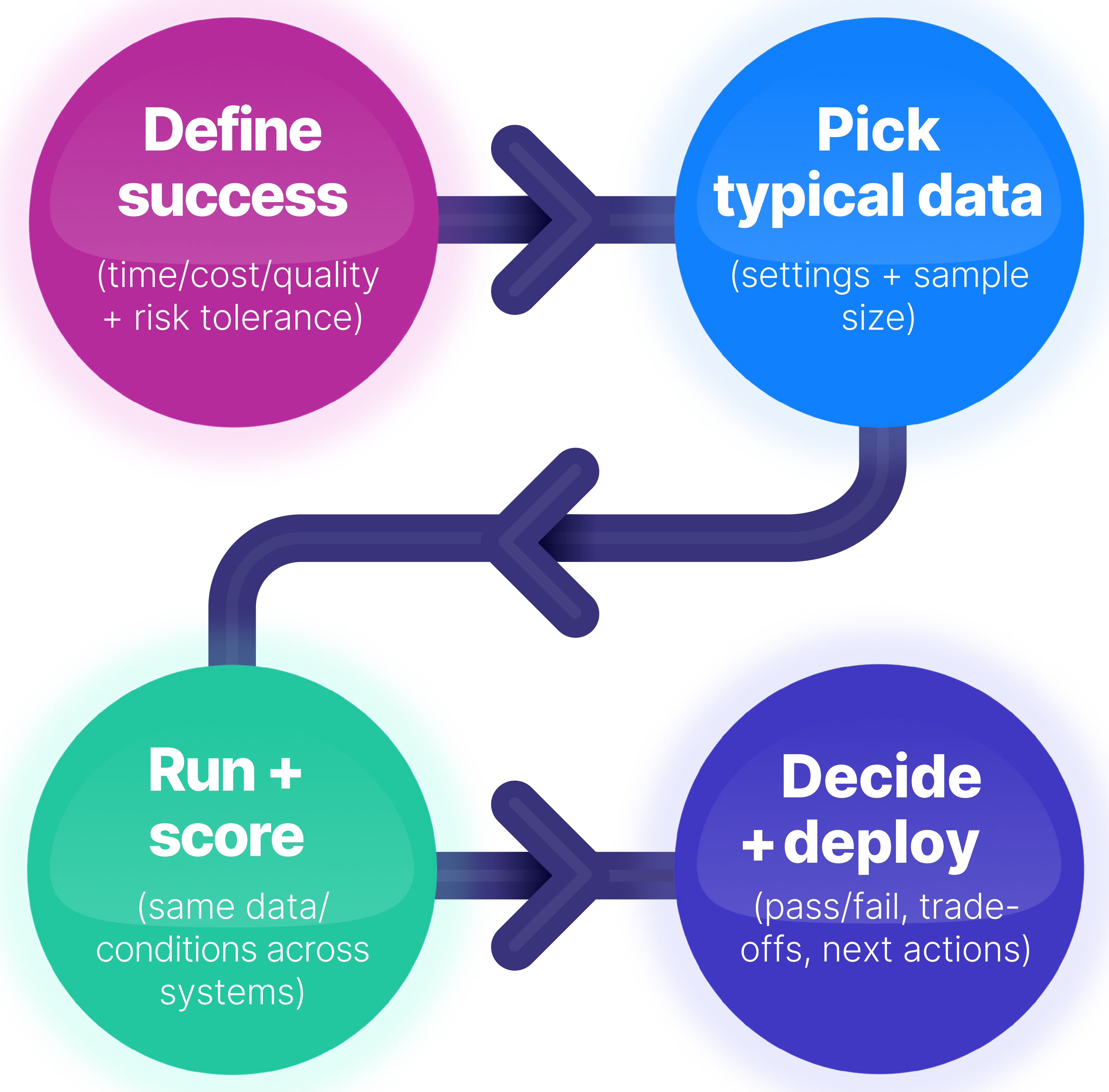
We recommend the following methodology for evaluating AI on your multilingual content, each step corresponds to a section below where we discuss recommendations in detail:

1 **Establish success criteria**
Understand your expectations in applying AI and establishing meaningful criteria that illustrate your tolerance for risk

2 **Build the right dataset**
Collect data to test that maximizes the utility of the evaluation

3 **Generate and evaluate the output**
Run your data through the solution and calculate your success metrics

4 **Interpret your results**
Understand whether the solution fits your needs and what options might be available to you



Define success
(time/cost/quality + risk tolerance)

Pick typical data
(settings + sample size)

Run + score
(same data/ conditions across systems)

Decide + deploy
(pass/fail, trade-offs, next actions)

1 Establish success criteria

In this section we'll recommend a framework for establishing meaningful success criteria that is grounded in understanding both your objectives in applying AI technology and your tolerance for risk.

By the end of this section you should be able to define a set of criteria that will form the basis of your evaluation.

The Phrase value framework

Whilst individual requirements will be extremely varied, most customers find that their expectations of AI solutions fall into three dimensions: Time, Cost and Quality. There is a natural tension between all three of these dimensions that requires compromise in solutions design:

Time:

A key benefit of AI is in reducing the turnaround time for multilingual content. Time in this case is also synonymous with scale. With the advent of generative AI, more content teams are under pressure to turn out more multilingual content in more markets as quickly as possible and currently this depends on the scaling of human effort to guarantee success. In extreme cases customers have real time expectations - in live scenarios such as closed captioning of multilingual speech, customers need the output immediately. In many cases time is a primary motivating factor in pursuing AI-based solutions, but even with AI in the mix not every solution produces instantaneous results.

Cost:

Another motivating factor in adopting AI-based solutions is the reduction in cost over human-based workflows. In most cases, even complex AI solutions that combine multiple tools are significantly more cost effective than human translators, editors and reviewers. There are a myriad of ways to optimize cost within Phrase including full automation, partial automation through intelligent content routing, Translation Memory leverage and content locking.

1 Establish success criteria

The Phrase value framework

Quality:

By far the most complex dimension; the quality of multilingual output means both the linguistic fidelity of the content and the degree to which it meets individual expectations. Quality can mean whether or not the language is fluent and grammatically correct, but also whether the content reflects the intent and style that a customer wanted. Quality expectations vary wildly between use cases; for some, getting the message across is enough, for others (especially in high risk or regulated use cases) quality is paramount.

Content outcomes and other expectations

For customers who are tied closely into content workstreams there are a multitude of other metrics at play. Engagement and conversion metrics might also be a consideration; if your multilingual content team is part of a broader content stream it's a great idea to understand how your internal customers are measuring success and where applicable, factor those metrics into your evaluation.

1 Establish success criteria

Defining objectives

For any evaluation, **understanding your own expectations per use case is absolutely critical.** The challenge is that more often than not, expectations are complex and not easily measured or interpreted. For this reason we recommend defining objectives along the lines of the value dimensions outlined above.

Objectives are non-negotiable targets couched in absolute terms. In so far as is reasonably possible, determine and define your expectations of time, cost and quality in this manner:

Time:
Depending on the use case it might be more or less relevant to consider completion time at the job level or even segment level for near real time use cases. For example: "Completion time of a job should not exceed X".

Cost:
Depending on your current implementation this can be expressed in terms of a maximum spend (and alignment with a related budget), a reduction over current spend or cost per unit of text (most commonly cost per word). For example, "Cost per word should not exceed Y".

Quality:
Given that quality can be subjective, understanding your quality expectation is usually a multistep process. Different metrics and methodologies provide different perspectives that are all useful in defining expectations:

1 Establish success criteria

Defining objectives

1 ***Linguistic quality: What degree of linguistic fidelity are you aiming for?***

MQM (either manually applied through human LQA or automatically using Auto LQA) can provide granular information on the nature of errors. One dimension that might be of relevance is error criticality. With MQM we might capture the number of errors considered ‘critical’, framed as “Jobs contain no critical errors” for example.

Metrics such as COMET and BLEU can be used in the context of comparison of systems, but these require that human translations exist for the evaluated content which isn’t always realistic at scale.

At Phrase we use QPS (Quality Performance Score) as a primary estimator of linguistic quality at scale. To simplify things further, we provide suggested bands of quality to aid interpretation of the output scores. An example might be “Jobs should be rated as at least “Excellent” according to QPS”.

Post edit metrics such as TER (Translation Edit Rate) can be used as a proxy for quality but this is not generally recommended as the volume of editing doesn’t correlate well with the actual quality. It isn’t possible for example to understand how critical an edit was to the translation.

2 ***Stylistic and functional expectation: Was the translation fit for purpose?***

Depending on the use case and the impact of the content in question, this may be more or less difficult to pin down. Metrics like QPS don’t indicate adherence to terminology or stylistic preference. They also don’t tell you whether the inline tags (for example HTML) were correctly positioned or whether consistency across sentences has been maintained.

For use cases where the focus of the solution is transformation of the content and (and likely deviation from the source, like Auto Adapt), looking at the linguistic fidelity is only partly useful. In this case we might consider more abstract criteria, such as “Content is consistent with the company brand and style guidelines”.

Some criteria can be automatically validated using QA checks (such as tag fidelity, segment length) or with MQM through human evaluation or in some cases (for example terminology adherence) with Auto LQA.

For a detailed explanation of which metrics to choose and when see the section on Metrics at the end of this handbook.

1 Establish success criteria

Example case study (Part 1):

In the following examples, we'll use a single fictional company, Acme Inc., to demonstrate each step of the evaluation process.

Acme Inc. wants to validate whether adding an Automated Post-Editing AI (APE) to their solution for website localization adds value. They currently use basic neural MT without bells and whistles but they've been struggling with output quality issues of English-Japanese.

For their evaluation they define the following expectations as objectives:

Time:
Jobs should take no longer than 5 minutes to complete

Cost:
Cost per word should not exceed $0.02

Quality:
Output should be 'Very Good' quality according to QPS

This gives Acme Inc. a baseline expectation for their content on which they can design their evaluation of APE and any other solution they might consider.

1 Establish success criteria

Establishing risk tolerance

As we highlight above, the goal of any evaluation of AI is to understand to what degree the performance of the systems matches our tolerance for risk. No AI solution will consistently meet your expectations 100% of the time, just as no human will either. Determining whether or not a solution is a good fit for your use case requires definition of your tolerance for failure.

Given your expectations on each of the dimensions defined above, what degree of failure are you willing to accept? Defining expectations as absolute is useful because it allows us to define risk as success criteria. Whilst objectives are absolute targets, success criteria define how often those targets are met.

To use an example above, "Jobs should be rated as at least "Very Good" according to QPS"; it is highly unlikely that any AI system will succeed in meeting this expectation 100% of the time in all languages. For this reason we determine a success criteria or threshold that illustrates our tolerance for risk. We might be accepting of say 10% of jobs falling short of this expectation, in which case we can define a success criteria for our evaluation as "90% of jobs should be rated as at least "Very Good" according to QPS"

In many cases it might be that the absolute objective reflects zero tolerance for risk. This is likely the case on the cost dimension, especially if the success criteria is implied by a budget constraint.

90%

of jobs should be rated as at least "Very Good" according to QPS"

1 Establish success criteria

Example case study (Part 2):

Acme Inc. defined their objectives for evaluating their AI as a solution for their website localization workflow. As a reminder, their objectives were defined as follows:

Time:
Job should take no longer than 5 minutes to complete

Cost:
Cost per word should not exceed $0.02

Quality:
Output should be 'Very Good' quality according to QPS

Given that their websites are static, it doesn't really matter to them that all jobs complete under 5 minutes, this is just a preference goal. They don't have flexibility in their budget and their objective is tied to total project cost so their tolerance for exceeding this is zero. They do want a majority of content to be of 'Very Good' quality but they also recognize the trade off in automating with AI and not using human translators so they are willing to tolerate some small portion of the content falling below this objective. In light of the above they define their success criteria as follows:

Time:
60% of jobs take no longer than 5 minutes to complete

Cost:
Cost per word does not exceed $0.02

Quality:
90% of content meets or exceeds 'Very Good' quality according to QPS

2 Build the right dataset

For an evaluation to be useful, we need to establish reasonable confidence that any conclusions we draw will be broadly applicable to any content the AI might touch in the future. As we've highlighted above, AI (especially generative AI) can be highly unpredictable and might work better for some content or languages than others. When we deploy production AI at Phrase, we evaluate it on as much and as many types of data as possible, but guaranteeing success for everyone is impossible.

The ideal dataset for evaluation will look different for everyone and depends on lots of things; the complexity of your data, the range of languages, the uniqueness of your content etc. The primary goal is to choose a 'representative' sample of data; testing on all of your data is impractical, but we can choose samples that give a sense of the big picture.

Before you start sampling data, it's important to understand the distribution of your content as best you can:

What domains (medical, legal, marketing) does your content exist in?

Do you have clear content types (customer service chat, blog posts, emails) that can be identified?

What languages/language pairs are you working with?

It is a good idea to map these out in a table or list to get a good idea of how broad your evaluation ideally needs to be.

2 Build the right dataset

Example case study (Part 3):

Acme Inc. maps out their content as follows:

Language Pairs:
English-Japanese
English-French
English-Spanish

Domains:
Marketing

Content Types:
Website

This gives Acme Inc. 3×1×1=3 evaluation settings to consider.

2 Build the right dataset

Here are some things to consider when choosing data for your evaluation:

Size of the dataset:
It's important to test on as much data as possible, remembering that you're going to need to collect some metrics on each datapoint, it isn't practical to consider millions of segments, but it is important to evaluate enough. What constitutes 'enough' depends on how diverse your data is; if your content is generally predictable and formulaic with lots of repetition you might be ok with a smaller sample. If your content is highly diverse with very varied vocabulary, you will likely need a lot more.

As a rule of thumb, you should ideally try for at least 1000 sentences per setting; say you have 3 language pairs, you'd ideally look to collect 3000 sample sentences to test.

In many cases, that won't be practical, either because running the evaluation on that volume is inhibitive or because you have many more than three settings to choose from. In that case you have a few options: You can either be selective with your settings (as described below) or reduce the sample size. Compromising on the size of the dataset isn't ideal but it can still give you a meaningful signal.

Variety of settings:
Most larger, enterprise teams will be working with a multitude of settings with many language pairs across many content types. This makes it necessary to be selective about which settings to test on. If you're lucky enough to have 3 settings to test on (like our Acme Inc. example above) then you should aim to cover all settings equally. If the breadth of your multilingual operations leaves you overwhelmed by the number of settings to test, focus on impact:

Identify the settings in which you have the most volume, the biggest potential engagement or the highest risk and prioritize these for evaluation. Given the impracticality of evaluating every setting at scale you will always have blind spots. Choose where you prefer these blind spots to be.

It is sometimes preferable to break the rule of thumb on volume established in the section above and prefer smaller samples for lower impact settings than to exclude those settings altogether.

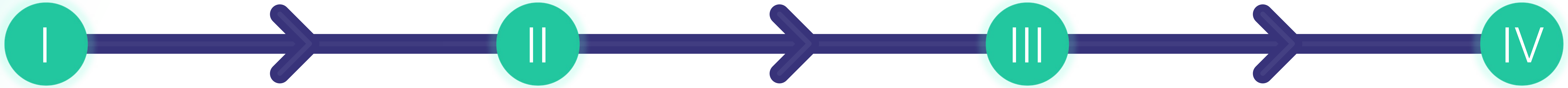
Diversity of the data:
We want the evaluation result to be broadly generalizable to the big picture. For this reason it is important that the evaluation data closely resembles the larger data distribution as much as possible. That said, it is common that data samples will contain repetition, either full or partial. If your content is formulaic in nature you may find that you have sample sentences that look very similar. If that reflects the nature of your content that's fine but it might produce a blind spot on performance on more diverse content. For this reason it is a good idea to check that your content is reasonably diverse. Repetitions are ok, but excessive repetition might impede the utility of your result.

⚠ **Tip:** Check the word count of the sentences in your data sample. If you have a good mix of sentences of different lengths you've probably got a decent sample of content.

3 Generate and evaluate the output

Once you've established your objectives and curated a representative dataset, the next step is to generate outputs from the AI system or combined solution and evaluate them against your chosen criteria.

Running your evaluation



Process the dataset

Run your sample data through the AI solution(s) you are testing.

Note

If you are comparing multiple systems, it is critical that you use the same dataset and conditions for each system so the results are comparable. Comparisons must be 'apples to apples'.

Capture outputs

Store AI outputs in a structured way so that they can be aligned with your evaluation framework (e.g., by language pair, content type, or domain). Using something like a spreadsheet is useful in this scenario; you can align the input and output text in rows with any corresponding metrics that you capture.

Apply metrics

Use the appropriate quality and efficiency metrics you identified earlier. Some of these may be easily captured through Phrase Data. Aligning these metrics at a sentence/ job level with the inputs and outputs can aid interpretation.

Resolve success criteria

Once you have the output metrics, for each system you've evaluated, review your success criteria and calculate success rates/metrics. This might include establishing another column with a label. For example you might look at your QPS score and (in a second column) use a formula to determine if the score was above a threshold defined in your success criteria. Determining the success rate in this example might involve calculating the number of samples that successfully exceeded this threshold.

3 Generate and evaluate the output

Example case study (Part 4)

Acme Inc. runs their 3-language dataset through both their baseline MT system and their MT system combined with their APE solution.

As a reminder they established the following success criteria:

Time:
60% of jobs take no longer than 5 minutes to complete

Cost:
Cost per word does not exceed $0.02

Quality:
90% of content meets or exceeds ‘Very Good’ quality according to QPS

They extract all of the relevant metrics after running both systems through identical settings and datasets and record the following success metrics:

English-Japanese	MT Only	MT+APE
Time: (% jobs < 5 minutes)	100%	96%
Cost: Average CPW	$0.0005	$0.01
Quality: (% jobs > “V Good”)	62%	92%

English-French	MT Only	MT+APE
Time: (% jobs < 5 minutes)	100%	98%
Cost: Average CPW	$0.0005	$0.01
Quality: (% jobs > “V Good”)	74%	93%

English-Spanish	MT Only	MT+APE
Time: (% jobs < 5 minutes)	100%	97%
Cost: Average CPW	$0.0005	$0.01
Quality: (% jobs > “V Good”)	79%	96%

This gives Acme Inc. a quantifiable basis for comparing the solutions.

4 Interpret your results

Interpreting results is not just about reviewing raw numbers but about understanding how they map back to your objectives, tolerance for risk, and overall business needs.

One of the reasons that establishing success criteria is important is that evaluations inevitably produce lots of metrics and likely a confusing picture. In some cases your system comparison (if that's the purpose of your evaluation) will produce a clear winner. In many cases however it will not. Following the framework above significantly reduces the complexity and clarifies decision making.

Compare against expectations

If you were able to establish success criteria you're likely able to determine "pass"/"fail" for each of your criteria. In an evaluation of a single system, determining that all criteria are met should give you the signal you need to make a decision. In the case of a system comparison, assuming both meet all criteria, ranking the systems on their success metrics and counting win rates might be an easy way to understand which system to prefer.

Near misses and failures

In most cases, if you're clear in your expectation and have set success criteria accordingly, interpreting results should be fairly straightforward. Sometimes however, you might be faced with an unexpected result. Imagine a scenario where the quality of a solution greatly exceeds your expectations but cost was slightly higher than you anticipated. At this point you might want to review the section above on Risk Tolerance and decide whether you want to revisit your success criteria to take advantage of new information revealed in your evaluation.

In some cases, failure might introduce a 'dead end' scenario that you weren't expecting. Imagine that you are comparing two solutions and neither solution succeeds in meeting your expectations. In this case it may be worth reviewing your criteria, it might be that your expectations exceeded the realistic capabilities of the AI solutions.

4 Interpret your results

Identify strengths and weaknesses

Defining success criteria along the three dimensions of the Phrase Value Framework will allow you to map where the system performance meets expectations and identify meaningful trade-offs. It is highly likely that performance will not be consistent across settings and you will find some settings on which your AI solution will work better than others. This can be extremely useful in determining how to apply the given solution:

Say for example you have a fixed budget for localization across multiple settings but you find out that quality of the AI is much higher than your expectation in a few settings; these would be great candidates for more automation.

On the other hand, say you find one solution to fall short, you might consider automating in most settings but retaining human-in-the-loop for the weakest setting.

Considering time, cost and quality - and weighing the tradeoffs between them - can be a great way to optimize the value of your AI solutions and understand where they fit your needs best.

4 Interpret your results

Example case study (Part 5)

Acme Inc. interprets their evaluation of the two solutions by reviewing the success metrics and determining where eithevr solution has been successful:

English-Japanese	MT Only	MT+APE
Time: (% jobs < 5 minutes)	100%	98%
Cost: Average CPW	$0.0005	$0.01
Quality: (% jobs > "V Good")	62%	92%

English-French	MT Only	MT+APE
Time: (% jobs < 5 minutes)	100%	98%
Cost: Average CPW	$0.0005	$0.01
Quality: (% jobs > "V Good")	74%	93%

English-Spanish	MT Only	MT+APE
Time: (% jobs < 5 minutes)	100%	97%
Cost: Average CPW	$0.0005	$0.01
Quality: (% jobs > "V Good")	79%	96%

Acme notes that the MT solution fell short of their quality expectation in all settings. Adding APE however met or exceeded all of their expectations making it a successful candidate for their chosen solution.

Acme notes however that the Quality in English-Japanese is determined to be at exactly 90%. They decide to keep an eye on this setting and rerun an evaluation at a later date to examine whether performance remains satisfactory.

4 Metrics

The following tables illustrate a variety of metrics that you might consider in your evaluation, each is described with reference to the Phrase Value Framework described above.

Time

Metric	What it Measures	Available in Phrase Data
Turnaround Time	End-to-end time to complete a job. Critical for scale and deadlines.	Yes, currently you can track the start time of a job in TMS.
Latency	Response time for individual translations (useful in realtime use cases like captions).	Limited – depends on workflow setup.

Cost

Metric	What it Measures	Available in Phrase Data
Cost per Word	Direct cost efficiency relative to volume of text processed on the platform.	Yes.
Total cost	Aggregate human and platform costs.	Yes - although related human costs may be tracked externally to the platform

Quality

Metric	What it Measures	Available in Phrase Data
QPS (Quality Performance Score)	Overall linguistic quality; combines fluency, grammar, adequacy, and severity of errors.	Yes
MQM (Multidimensional Quality Metrics)	Detailed error categorization (critical, major, minor) across lexical, semantic, stylistic, and functional dimensions.	Yes (via human LQA or Auto LQA)
Post-Edit Distance / Translation Error Rate (PED / TER)	Degree of editing required by humans (irrespective of error severity)	Yes
BLEU (Bilingual Evaluation Understudy)	Lexical overlap with reference translations (n-gram similarity). Sensitive to surface word choices.	Limited – available through CustomAI
COMET	Semantic similarity and adequacy of translation vs. reference, using neural models.	Limited – not directly in Phrase Data, but can be computed externally or via CustomAI
Tag Fidelity / Formatting Accuracy	Correct handling of inline tags, placeholders, or formatting tokens.	Yes (via QA checks)
Terminology Adherence	Compliance with term bases, glossaries, and brand lexicons.	Yes (via Auto LQA / QA checks)

About Phrase

Phrase is the world's leading enterprise AI-led language technology platform, helping global businesses scale with confidence, consistency, and speed. It empowers teams to adapt products, content, and customer experiences for every market, turning localization into a growth engine rather than a bottleneck.

From market entry to brand expansion, Phrase provides the infrastructure that enables organizations to launch, communicate, and operate effectively across languages and regions. Its unified platform connects people, processes, and AI, streamlining everything from product updates and marketing campaigns to customer support and documentation.

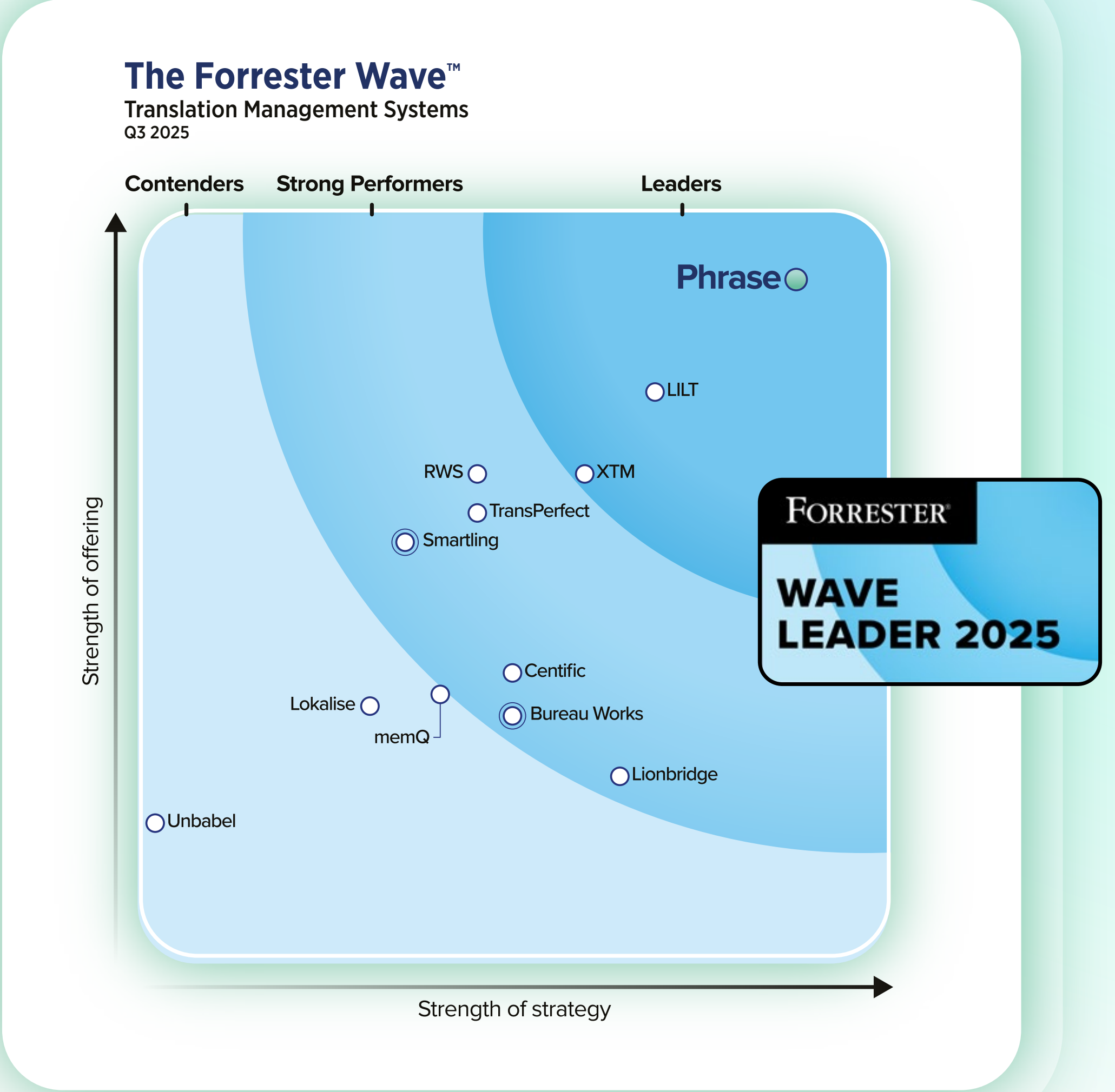
By combining automation, machine translation, and AI-powered quality management with human expertise, Phrase gives enterprises full control over quality, cost, and time-to-market. Teams can execute multilingual launches, maintain brand integrity, and ensure compliance in one secure, scalable platform.

Recognized as a Leader in The Forrester Wave™ for Language Services Technology and rated a top platform on G2, Phrase supports the world's most ambitious companies in building stronger, more consistent global operations.

With enterprise-grade security, integrations across tech stacks, and a proven track record in global expansion, Phrase helps organizations grow internationally, efficiently, sustainably, and without compromise.

Read the full Forrester report

We're a Leader



Top scores in 21 out of 26 categories

Customers love Phrase's practical, AI-forward thinking

Thank you

Phrase

phrase.com

aws PARTNER Retail Software Competency

FORRESTER WAVE LEADER 2025