LARA 3

Overview & MT Capability Comparison

By Kirti Vashee

translated.

Executive Summary: Introducing Lara 3

Lara 3 marks a foundational paradigm shift for Translated, moving away from conventional imitation-based machine translation toward a training methodology centered on "learning by experience." This approach allows the model to internalize knowledge autonomously via practice runs and structured feedback loops, escaping the strict limitations of merely reproducing static translation memory records. Consequently, this generation delivers significant quality advancements within a robust, enterprise-grade framework.
According to Translated, this practice-based environment has driven the most substantial quality leap between successive generations in the system's history. These performance enhancements were validated through double-blind evaluations conducted by professional translators, who assessed Lara 3 output against platforms including Google, DeepL, and current general-purpose AI models.

Beyond core quality upgrades, this release introduces multimodal capabilities spanning image and audio translation alongside advanced developer and enterprise tooling. Structurally, it updates commercial engagement by replacing legacy per-seat licensing with an enterprise-wide operating model tailored specifically to the distributed workflows typical of modern global organizations.

According to Translated, this practice-based environment has driven the most substantial quality leap between successive generations in the system’s history. These performance enhancements were validated through double-blind evaluations conducted by professional translators, who assessed Lara 3 output against platforms including Google, DeepL, and current general-purpose AI models.

What Is Lara, and Where Does It Come From?

Lara is the LLM-based translation AI developed by Translated, a company that has been building adaptive machine translation technology since it launched ModernMT in 2016. Translated was named a Leader in the IDC MarketScape Worldwide Machine Translation Software 2022 Vendor Assessment, assessed ahead of Google, Amazon, and Microsoft in that quadrant, with IDC citing "continuous innovation and a strong multi-domain model" as the primary differentiators. CSA Research also recognized Translated's adaptive MT as "the most advanced implementation of responsive MT for enterprise use" in its 2023 Vendor Briefing.
Every major advance in machine translation's history has followed a shift in learning method: from hand-written rules (RBMT), to statistical examples drawn from parallel corpora (SMT), to neural end-to-end learning (NMT), to the self-attention mechanisms of transformers, and then to LLM pretraining.

Lara 3 introduces what Translated calls the next step in this technological progression: learning from experience. Translated frames this as a shift from "learning by imitation" to "learning by doing," tied to what its research group, Imminent, calls the "Age of Experience" in AI.

Independent researchers have moved in a similar direction: Reinforcement-based approaches to machine translation, optimizing models against quality signals rather than static reference data alone, have been an active research area since at least 2023, when early studies showed measurable gains from incorporating human-feedback-derived reward signals into MT training.

What Changed from Lara 2 to Lara 3

- **The Core Innovation—Learning by Doing: From Imitation to Experience**
- **A Broader Capability Set: Multimodal Translation and Language Asset Management**
- **Updated Enterprise Operating Model**
- **Deployment Surface and Integration Capabilities**
- **Introduction of Lara Think and Lara Prosa**

The Core Technical Shift: Imitation to Experience

Previous versions of Lara, like most machine-translation systems, learned by imitation:

The model was shown a professional translation (TM) for each source sentence and trained to reproduce it as closely as possible. That method has a ceiling because any sentence can be translated correctly in more than one way, so a model trained to match

a single reference learns to copy an answer without learning what makes it good.

Lara 3 breaks this constraint through a three-stage practice environment built into the final phase of model training.

1. Explore: For each training sentence, the model generates its own alternative translations, testing different terminology, phrasing, and stylistic choices—not copying a reference.
2. Get reviewed: An automatic judge evaluates every alternative and provides structured feedback on what works, what doesn't, and why. Critically, this judgment was built on Translated's documented expertise in how professional reviewers actually evaluate translations—their correction patterns, terminology standards,

   and reasoning.
3. Refine: Lara reinforces choices that earned positive feedback and corrects those that didn't, repeated millions of times across the training corpus.

Translated frames this iterative environment—practice and explore, review, and refine—as a mirroring of how professional linguists develop expert judgment, transitioning from reproducing historical translation memory examples to internalizing the foundational principles of contextual accuracy.
Translated has separately described this reinforcement-learning approach as training Lara to generate translations that are more likely to be approved by human experts, thereby reducing post-editing effort and raising overall quality. This approach mirrors how human translators improve through practice and reviewer feedback rather than by memorizing model answers.
For procurement teams, the practical takeaway is that Lara 3 is not a cosmetic update. It represents a structural shift in how the model is built and trained, which typically yields more durable quality gains than incremental prompt or feature tweaks, which are most frequently used by competitive alternatives.

A Broader Capability Set: Multimodal, Multi-Surface, Programmable

Lara 3 extends beyond text into image and audio translation, and product access expands across web, mobile, browser, and developer channels.

- **Image translation**: Three modes ranging from simple text overlay to full generative recreation of the image around the translated content, useful for marketing and brand assets where layout and typography carry meaning.

- **Audio translation**: Audio-to-audio (STS) and audio-to-text (STT), currently built for single-speaker content; multi-speaker support with voice cloning is planned for later in 2026.

- **Document handling**: Fewer formatting and layout errors, batch translation of many files into many languages with a single ZIP download, and a toggle to translate text embedded inside images within PDFs.

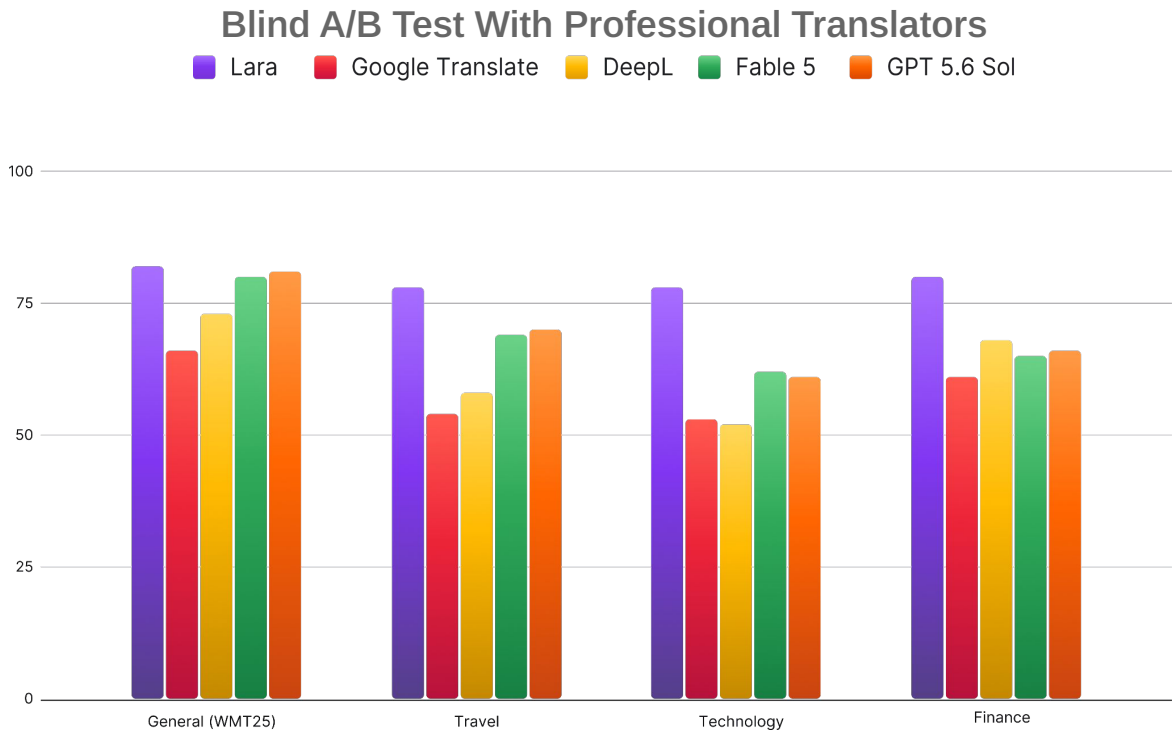
- **Platform reach**: Native iOS and Android apps, a Chrome and Edge browser extension, a Google Sheets integration, and a CLI for CI/CD pipelines.

- **Developer access**: Nearly all new capabilities, including image and audio translation, are available through Lara's API and across its SDK (Python, JavaScript/TypeScript, Java, PHP, Go, .NET, Swift), plus an MCP server connecting Lara to AI assistants including Claude, ChatGPT, and Grok.

- **Ecosystem**: Connections to the CAT tools and platforms that localization teams already run, incl

July 2026 Comparative Evaluation of Lara 3 vs. Leading Market Alternatives

Translated validated Lara 3 through blind human evaluation using data from WMT 2025*, a public benchmark spanning books, news, and user conversations. Three target-native professional translators reviewed anonymized outputs from Lara 3, Google Translate, DeepL, Fable 5, and GPT-5.6 Sol side by side. Lara 3 achieved the highest majority-approval rate among the systems tested.
The enterprise benchmarks extend this general-domain baseline to travel, technology, and finance, where Lara 3 ranked first in every domain, with double-digit leads in technology and finance and a clear lead in travel.



Comparative System Output Evaluation Process

Selection of professional translators
To assess translation quality, we selected top-performing professional translators from a network of 700,000 using T-Rank, an AI-driven ranking system developed by Translated. T-Rank helps to select top-performing, domain-qualified professional translators by evaluating their past performance and expertise across more than 30 criteria. This ensured that the translators selected for evaluation were highly qualified native speakers of the target languages.

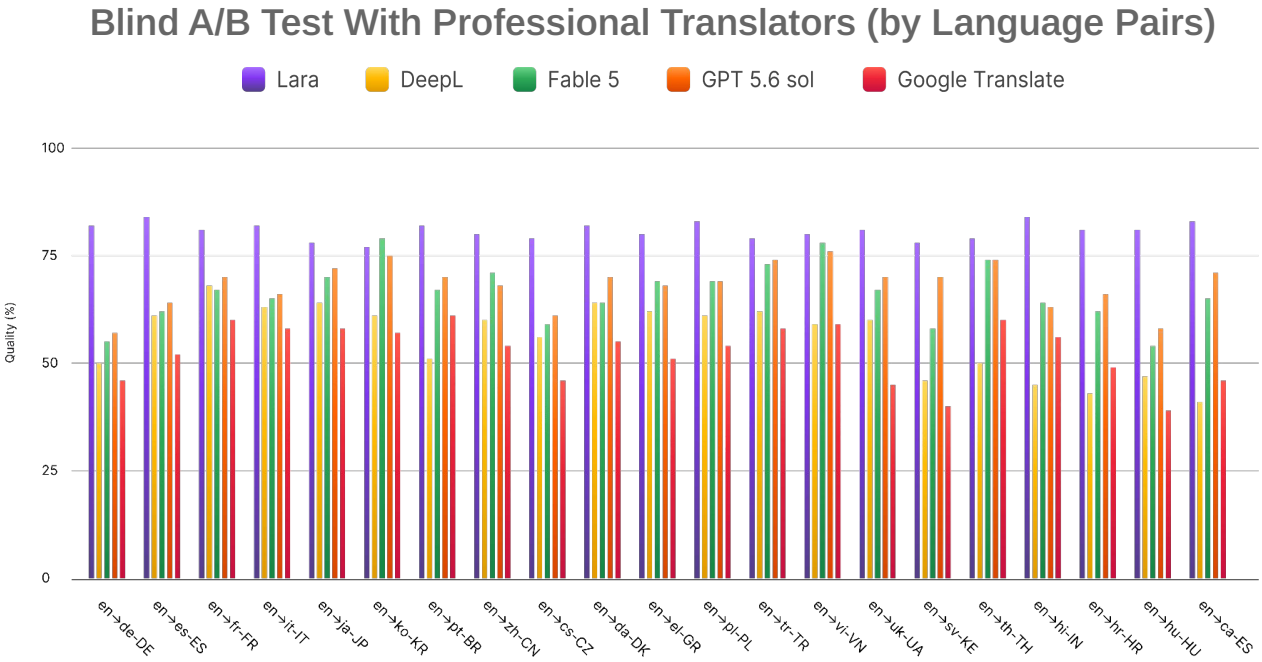
Human evaluation
Three target-native professional translators independently reviewed each item. Each reviewer saw the source, available context, and the anonymized outputs of all five systems side by side, without seeing system identities, or the other reviewers' judgments.

*Translated has published all Lara 3 translation outputs from its WMT 2025 evaluation, together with the evaluation methodology and validation documentation, on GitHub. This allows organizations to inspect Lara 3's outputs, assess the approach, and apply the same evaluation protocol to their own content.

Majority agreement
Each reviewer assigned a separate Good/Bad judgment to every anonymized system output. A system output received majority approval for an item when at least two of three reviewers marked it Good for the declared professional use case. More than one system, or none, could receive approval for the same item. This method reduced subjectivity and emphasized consensus.

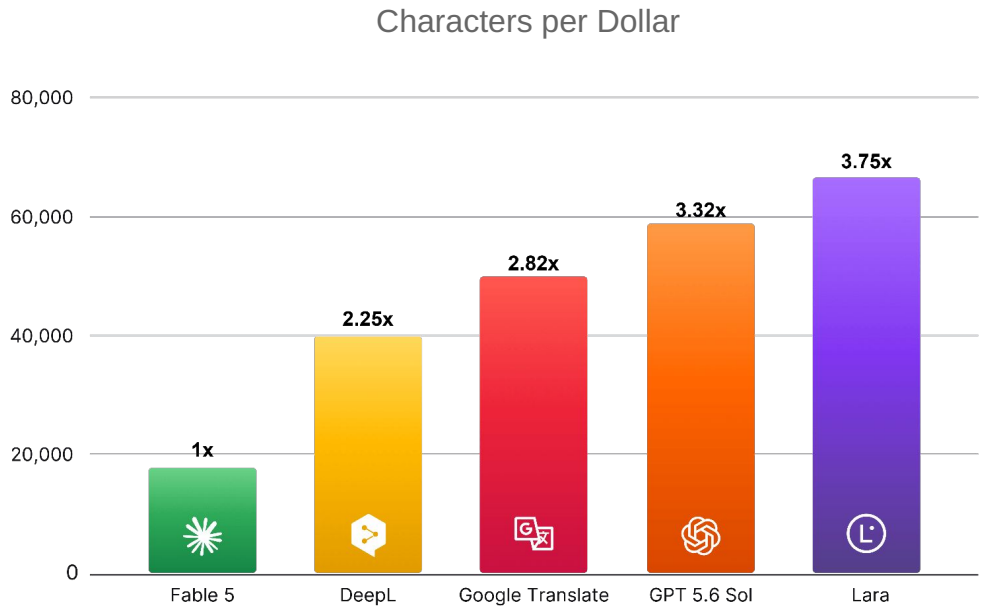
Scoring methodology
The reported majority-approval rate for each engine is the percentage of evaluated source-target items for which at least two of three reviewers marked that engine's output Good.

We designed this evaluation to compare the performance of various machine translation engines using real-world, enterprise-level content. Our test set consisted of about 16,800 sentences, consisting of 200 English source sentences translated by machine-translation systems into twenty-one of the most frequently requested localization languages, as well as some less-used languages. The accuracy of these machine-generated translations was meticulously assessed by professional translators carefully selected for the review process. To ensure objectivity and eliminate bias, we employed a double-blind method: The reviewers were unaware of which machine translation engine produced each translation, and they were not informed of other reviewers' evaluations. This approach allowed for an unbiased and fair assessment of each system's performance.



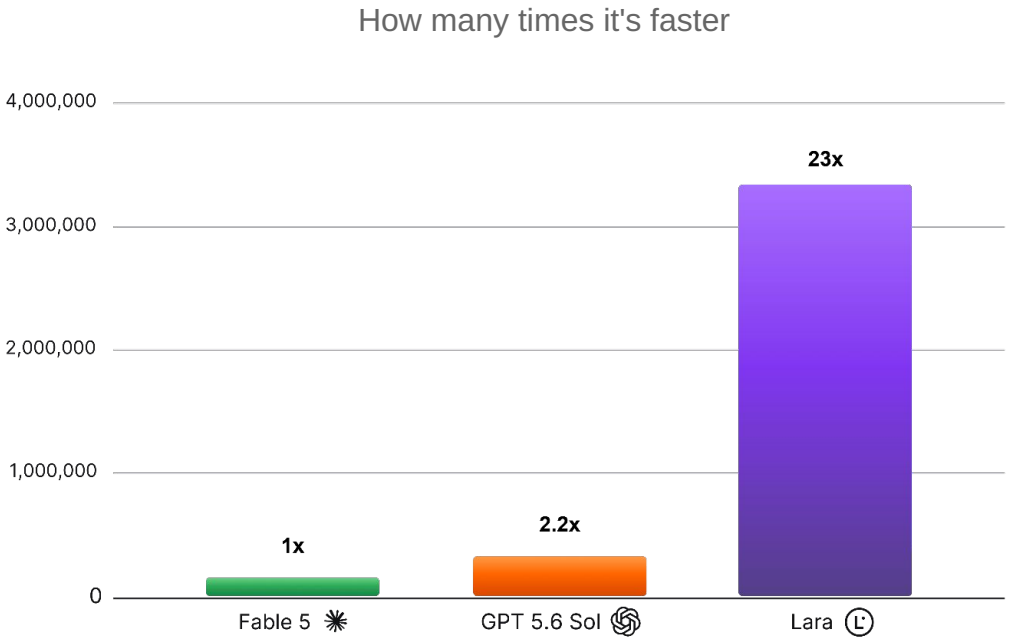
The MT system performance profile is equally important for procurement decisions. In production environments where both throughput and unit economics matter, this is a significant operational advantage. The only evaluated system that exceeded Lara on raw processing speed ranked last on quality.

For Cost Efficiency, Higher is better: Lara is almost 4X MORE Efficient than Fable 5

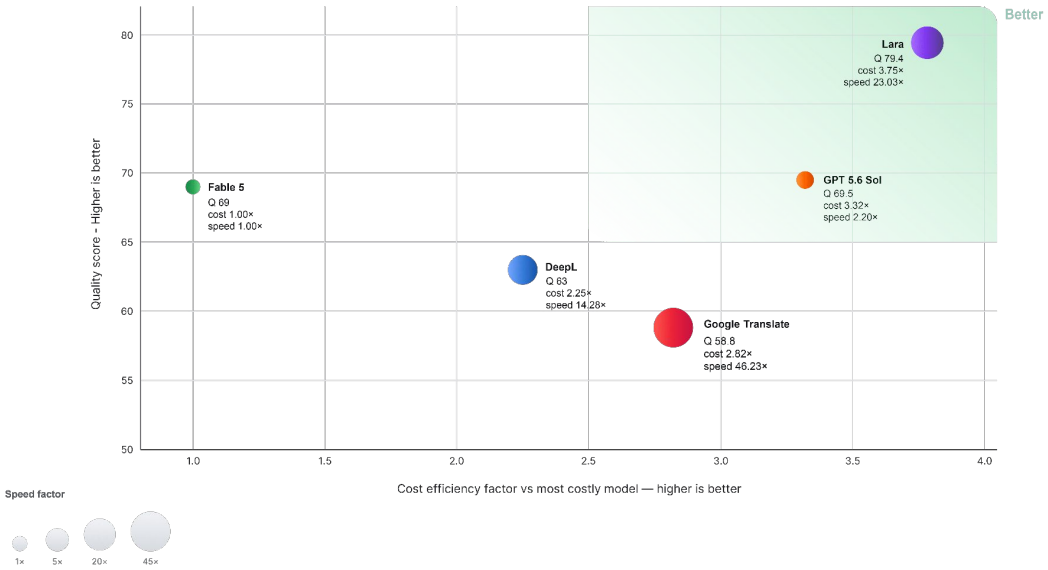


When contrasted with the highest-priced model under evaluation (Fable 5), Lara 3 provides superior quality at approximately a quarter (25%) of the cost and at over twenty-three times (23x) the speed.

Lara is the fastest LLM-based solution available (Fable 5 = 1)



Quality vs Cost Efficiency - Speed Improvement as a Bubble Size



This chart plots quality on the y-axis and cost-efficiency relative to Fable 5 on the x-axis; bubble area represents the relative throughput factor.

Compared against Fable 5, the most expensive system in its evaluation set, Translated reports the following:

Dimension	Reported Lara Results vs. Fable 5
Quality	Highest-rated output across all domains tested
Cost	About 3.7 times more cost-efficient
Speed	About 23 times faster

For procurement and IT teams, this combination matters more than any single metric. Machine-translation deployments are usually constrained by whichever quality, cost, or speed is weakest; a model that trails on any one dimension often gets abandoned in production even when it is strong elsewhere. Translated's position is that Lara 3 avoids that trade-off across all three.

Governance and Advanced Reasoning Modes

- Multilingual, multidirectional glossaries and translation memories can now be managed directly in the web app, SDK, or API, and apply consistently across the entire Lara ecosystem.

- Profanity detection and filtering are now built into the translation API, letting compliance teams flag, mask, or block offensive content in source or target text.

- **Lara Think** is a reasoning model that runs multi-step analysis against glossaries and style guides, catching up to 80 percent of major terminology and style issues before human review, at the cost of longer processing time and slower throughput.

- **Lara Prosa** applies an explicit editorial style guide to preserve tone, rhythm, and voice for content where brand identity matters, such as marketing and literary material.

A reimagined Enterprise Commercial

The Slator 2025 Localization Buyer Survey identified operational and integration complexity among the most persistent barriers to realizing AI value in localization. A significant part of that complexity is structural: Most enterprise translation tools were designed for localization departments, not for the distributed translation reality of modern global organizations. Sales teams translate proposals. Legal reviews clauses in foreign-language contracts. HR distributes policy documents. Product teams localize release notes. Marketing adapts campaigns. These needs are distributed, often unpredictable in timing, and incompatible with per-seat licensing models that require planning of who will need access.

Translated is replacing traditional, seat-based licensing with a model built around three observed enterprise realities: Translation demand is distributed across departments, arrives in unpredictable waves, and requires consistent brand voice across teams. Translated is replacing per-seat licensing with three changes:

● Unlimited seats: Access is no longer capped by license count, on the premise that translation needs in large organizations are distributed across sales, legal, HR, support, and marketing rather than confined to a localization team.

● Overquota: Usage above a committed baseline is not blocked; teams can keep translating through demand spikes such as product launches or RFPs, with commercial reconciliation handled separately from the workflow.

● Role-based asset control: Only designated users can edit glossaries and translation memories, but the resulting terminology and brand rules apply automatically to every user across the organization.

Translated frames this as a shift from selling translation seats to supporting translation operations enterprise-wide. For procurement, the practical implication is a different total-cost-of-ownership conversation: a baseline commitment plus overquota terms, rather than a per-user license count that has to be renegotiated every time it needs change.

This shift is consistent with the broader market direction identified by industry analysts, who note that language service and technology providers are moving from siloed localization tools toward unified, enterprise-wide content workflows, sometimes described as a "post-localization era" where translation becomes embedded infrastructure—like email—rather than a departmental purchase. For procurement and executive stakeholders, this model reduces the administrative overhead of seat forecasting and license reallocation while placing quality governance in the hands of a defined set of linguistic experts. This structure aligns with enterprise interest in centralized control paired with decentralized access.

Translated also disclosed an on-premises Blackwell GPU cluster, co-designed with NVIDIA and Lenovo, running alongside a hybrid architecture that draws on AWS for burst capacity and failover. The company positions this as the infrastructure basis for the quality, cost, and speed results described above, rather than a separate announcement.

Summary of Key Features of Lara 3

Dimension	What Lara 3 Delivers
Model quality	Highest human-rated quality in blind evaluation across Google, DeepL, Fable 5, and GPT 5.6 Sol
Cost efficiency	Approximately 3.7x more cost-efficient than Fable 5
Speed	23.03 × the measured throughput of Fable 5 in the declared single-thread benchmark.23x
Learning method	Experience-based (reinforcement from expert reviewer feedback), not TM imitation-only
Modalities	Text, documents, images (patch/inpainting/generative), audio (audio-to-audio and audio-to-text)
Advanced reasoning	Lara Think: Significant reduction in major terminology and style errors
Style fidelity	Lara Prosa: Editorial style guide layer for voice-sensitive content
Integration breadth	Browser, mobile, CLI, MCP, major CAT tools, localization platforms, CMS connectors
Vendor standing	Gartner Cool Vendor, IDC MarketScape Leader, Enterprise MT Software; CSA Research "Most Responsive MT"

Why This Matters for Buyers

For enterprises evaluating translation technology, the relevant question is rarely whether AI translation meets a quality threshold—leading systems have broadly reached that threshold. The more consequential questions concern total cost of ownership, governance of linguistic consistency across distributed teams, the availability of expert consultative support, the operational impact of constraints during demand peaks, and the organization's ability to extend translation to content types beyond plain text. Lara 3 addresses each of these dimensions in a single platform, built by a vendor with a documented track record in production translation at enterprise scale.

Two broader trends give this product release further context. Enterprise buyers are under growing pressure to show measurable return on generative AI investment rather than simply adopt it. CSA Research's 2026 industry predictions point to AI governance and differentiation, not automation alone, as factors that will separate organizations that scale global content well. And Gartner has projected that more than 80 percent of enterprises will have used generative AI APIs or deployed generative AI applications in production by 2026—up from under 5 percent in 2023—so most buyers evaluating Lara 3 already have an internal framework for judging GenAI vendors on more than raw model quality.

Separately, Forrester's Total Economic Impact research on comparable AI translation platforms has found ROI figures as high as 345 percent over three years, driven by faster translation cycles and reduced workload rather than headcount cuts, a pattern that reinforces the business case for adopting continuously improving MT engines like Lara.

For enterprises already running Lara 2 in production, the Lara 3 upgrade offers a low-friction quality lift since it deploys through the same TranslationOS, API, and CAT tool integrations already in use, including Trados Studio, memoQ, Crowdin, and Phrase. Enterprises evaluating this upgrade should weigh it against the total cost of switching versus staying, factoring in that the quality gains are backed by both human linguist review and statistically validated automatic scoring rather than single-metric benchmarks alone. Given the scale of the reported gains, particularly the up to 76-point improvement on user-generated content, Lara 3 represents a meaningful checkpoint for any global enterprise reassessing its MT vendor stack in 2026.

Lara 3 represents a well-documented, independently evaluated step forward—one built on a decade of Translated's investment in data quality, adaptive model architecture, and professional linguist collaboration.

When assessing Lara 3, enterprise decision-makers should focus on three core dimensions:

First, verifying if the reported quality, cost, and speed advantages hold during a pilot with your own content;

Second, determining whether the updated commercial framework aligns with your organization's actual translation distribution workflows;

And third, evaluating if the built-in governance features—including glossaries, asset roles, and profanity filtering—satisfy your specific compliance requirements. Each of these components can be tested independently before making a broader organizational commitment.

The combination of a foundationally improved model, governance controls, and continuous reinforcement learning addresses the specific concerns—accuracy, quality, and trust—that Slator identifies as the primary barriers to MT adoption in enterprise settings. For global organizations selecting a vendor, Lara 3 raises the performance floor across the entire language portfolio while preserving the adaptive, context-aware architecture that enterprise buyers have relied on since ModernMT's original introduction.

Pitfalls of Enterprise MT Quality Evaluation: Implementing a Robust Framework

Why the way you test and evaluate machine translation matters as much as which system you choose

Machine translation (MT) is now core business infrastructure for thousands of globally operating enterprises. Yet the methods most organizations use to select, compare, and govern their MT systems remain poorly suited to the realities of enterprise production environments. Evaluation errors made during procurement and deployment compound into downstream quality failures, cost overruns, and vendor lock-in.

Yet most enterprise buyers still evaluate MT the way researchers do: with a single benchmark score. **BCG's research on enterprise AI value creation, "Are You Generating Value from AI?" found that across hundreds of deployments, only around 10 percent of value achieved could be traced back to the model itself; the rest came from data, workflow integration, and human processing.** The same logic applies directly to MT selection. The evaluation method a procurement team chooses predicts downstream quality, cost, and vendor lock-in risk more reliably than the underlying model does.

The seven pitfalls below are the most common, and most expensive, mistakes enterprise buyers make when evaluating MT for production use.

1. Evaluating Generic Systems Instead of Adapted Ones

Most vendor demonstrations and third-party rankings compare stock engines in their out-of-the-box state. But enterprise MT performance is driven by domain adaptation, i.e., how well a system learns a company's terminology, brand voice, and subject matter, not by how it scores on generic news text.

The procurement question that matters is not "What did this system score on a public benchmark?" but **"How quickly will this system learn our content?"** A system that starts modestly but adapts fast could outperform one that scores well generically but never improves.

2. Testing on Small, Generic Samples that Produce Misleading Scores

Evaluations built on a small sample of generic sentences, or on samples that the vendor supplies rather than the buyer, produce scores that are statistically fragile and easy to game. **A credible enterprise evaluation requires a substantial, representative sample of the buyer's own content**, on the order of 1,000+ segments per language pair, reviewed by domain-qualified human evaluators. Smaller samples invite noise that gets mistaken as a meaningful signal.

3. Relying on Public Test Sets such as WMT and FLORES

Public benchmarks were built for the research community to compare general-purpose model progress, not to predict how a system will handle a bank's compliance disclosures or a medical device manufacturer's instructions for use. There is also growing evidence that public benchmark results are themselves unreliable.

Slator reported on 2025 research from Google and Boston University showing that even modest data contamination, test content that has leaked into a model's training data, can inflate BLEU scores by as much as 30 points in larger models.

Vendors—particularly major LLM providers—frequently test on data the model has already encountered during training. Buyers should treat public benchmark claims with the same scrutiny applied to a vendor-supplied ROI case, and insist on evaluation against test data the buyer controls and can confirm is unseen by the system.

4. Treating Quality Scores as Static Ratings

A score from six months ago describes a system that may no longer exist. Static models change with each new release, and adaptive systems—such as Translated's Lara—are built to keep improving as human linguists correct output. Scoring an adaptive system once and filing the result away misses the point of adaptive MT entirely and unfairly penalizes the systems designed to improve fastest on a buyer's own content. Best practice is to require time-stamped, version-pinned results, and where possible, to run a controlled adaptation trial measuring improvement over 30 to 60 days on the buyer's own content, not just performance on day one. The objective is to understand how quickly a system can be expected to improve over a defined period of standard production use. **The new question is not "Which engine scores highest?" but "Which solution, configured to our requirements, will produce the fewest errors for our content and use cases?"** A dynamic perspective has more value than a snapshot in time.

5. Misreading What Automated Metrics Actually Measure

BLEU, COMET, and MetricX were built for MT researchers to track iterative model improvement, a legitimate purpose in that context. They were not built to support a one-time vendor selection decision, and applying research-mode metrics to a buyer-mode decision is the single most common analytical error in enterprise MT procurement.

The scores have real blind spots: WMT24 evaluation found COMET-22's system-level correlation with human quality ratings was only around 0.69, and separate analysis has shown an empty translation can still score roughly 0.32 on COMET, while a wrong-language response can outscore a correct one.
Even Ricardo Rei, one of COMET's own creators, has stated publicly that no single metric is sufficient on its own. Automated scores are a useful triage signal. They are never a final verdict, and should always be triangulated against other measures.

6. Conflating Linguistic Quality with Business Fitness

A translation can be linguistically excellent and still fail the business. eCommerce listings, user-generated reviews, customer support tickets, and legal discovery documents each carry different quality bars. For many of these use cases, what matters is measurable business outcome, deflected support tickets, faster post-edit turnaround, and improved conversion, not incremental gains in a semantic similarity score.

Business fitness encompasses dimensions that no linguistic quality assessment captures: the system's integration depth with existing TMS and content platforms; the speed at which it improves when fed corrective feedback from in-house translators; its throughput and latency characteristics under production-scale volume; how well it enforces brand-specific terminology and style guides; and whether it supports the governance and audit requirements of regulated industries such as financial services, pharmaceuticals, or healthcare.

Forrester research on enterprise localization has consistently emphasized that organizations fail not because their MT output is linguistically imperfect, but because the selected system does not fit the enterprise's operational reality—creating bottlenecks in post-editing workflows, terminology inconsistencies across product lines, or compliance vulnerabilities in regulated content.

CSA Research's language risk framework, developed across a nine-report series, makes the same distinction from a risk-management perspective: The appropriate level of MT quality is not the highest achievable linguistic score but rather the level sufficient to avoid harm to customers, the brand, or the organization's legal standing.

Translated has proposed Time to Edit, the actual editing effort a professional linguist needs to bring MT output to publishable quality, as a more production-relevant measure than static, reference-based scores because it ties directly to cost and throughput. Enterprise frameworks should define quality tiers by use case, human-quality, light post-edit, or gist-only, before testing begins, rather than applying one linguistic bar to every content type.

7. Ignoring the "Illusion of Control" with General-Purpose LLMs

A growing number of enterprises are deploying general-purpose large language models—accessed via public APIs from major providers—as translation engines, either directly through prompt engineering or as the underlying layer of translation features embedded in enterprise software. This approach offers apparent simplicity: no specialized procurement, no linguistic asset management, no training cycles. It also creates a governance risk that most enterprise technology and procurement leaders have not yet fully recognized.

General-purpose LLMs are updated continuously and silently. The model a buyer accesses via API today is not guaranteed to be the same model they access tomorrow. Vendors reserve the right to update model weights, change decoding parameters, adjust safety filters, and retire model versions on schedules that are not necessarily disclosed to enterprise customers in advance. For translation-intensive workflows, where consistency, terminology precision, and regulatory accuracy are operational requirements, unannounced behavioral changes represent a material risk.

KPMG's research on AI risk management notes that 82% of enterprise leaders identify risk management as their primary AI governance challenge, and that the absence of model-versioning transparency and audit trails is a central concern for regulated industries. The Forrester Localization report similarly identifies "unsafe LLMs" and ungoverned AI as a leading source of localization failures, warning that deploying unbranded, unvetted general-purpose LLMs for enterprise translation multiplies risk across every language the organization operates in.

Calling a general-purpose LLM through an API can feel like control because the buyer decides when and how to use it. In practice, the vendor can update the underlying model at any time, often without notice, silently changing output behavior inside a system already running in production.

Gartner has found that roughly 30 percent of generative AI projects are abandoned after proof of concept due to inadequate risk controls, and Forrester's 2025 research shows enterprises shifting decisively toward buying governed AI solutions rather than building on raw model access, from 47 percent of organizations in 2024 to 76 percent in 2025.

For regulated industries such as legal, medical, and financial services, an undocumented behavior change in a production translation pipeline is not a minor inconvenience; it is a compliance exposure. Buyers should ask directly whether model updates are disclosed, whether a version can be pinned, and who owns the correction data used to keep the system improving.

Pitfall	What gets missed	Enterprise risk
Generic vs. adapted systems	Real-world domain performance and speed of adaptation	Wrong vendor selected for the content that matters
Small, generic test samples	Statistical reliability of the score itself	Decisions built on noise mistaken for signal
Public test sets (WMT, FLORES)	Contamination and production relevance	Inflated vendor benchmarks, false confidence
Scores treated as static	Continuous improvement of adaptive systems	Penalizes the best long-term vendor fit
Automated metrics misread	Gaps between metric score and human judgment	False confidence in a flawed single number
Linguistic quality vs. business fitness	Business outcomes the metric was never built to capture	Customer experience and brand consistency damage
Illusion of control with general LLMs	Undisclosed model updates in production	Regulatory, legal, and compliance exposure

The Bottom Line for Enterprise Decision-Makers

None of these pitfalls is hard to correct once named. The common thread is that most enterprise MT evaluation still borrows the habits of MT research: generic test sets, one-time scores, and a single automated number treated as ground truth. Enterprise procurement needs a different discipline: representative, buyer-owned test data; a hybrid metrics stack anchored by qualified human evaluation; an explicit test of adaptability over time, not just day-one quality; and governance terms that hold up after the vendor's next model update. Buyers who evaluate MT this way are not simply choosing a better score. They are protecting brand consistency, regulatory standing, and customer experience in every market the business operates in.

Why Model Selection Alone Is Not a Strategy

Enterprise AI conversations are dominated by model comparisons—which engine scores highest, which vendor claims the newest benchmark win—but this framing obscures where value actually gets created. BCG's research on enterprise AI deployments—"Are You Generating Value from AI?"—found that across hundreds of implementations, only about 10 percent of value achieved could be traced back to the model itself, 20 percent to the surrounding technology stack, and 70 percent to people and process.

The organizations that generated measurable value did not select the best-performing model—they invested in the 70 percent that no model, however capable, can provide on its own.

A useful discipline for evaluating any AI vendor claim, including MT vendor claims, is to ask what the speaker's business model rewards them for saying. Model vendors (OpenAI, Claude) have an incentive to frame model quality as decisive; tooling vendors (TMS, aggregators) have an incentive to frame flexibility and switching as the priority; infrastructure providers (Google) have an incentive to frame everything as commoditized compute. None of these framings is dishonest, but each describes the speaker's revenue model as much as it describes what drives enterprise value. Buyers evaluating machine translation should apply the same scrutiny to comparative benchmark claims, including the ones in this document.

Applied to MT procurement, engine choice is a real decision but a comparatively small one; the larger determinant of business value is whether the organization builds the workflow integration, human feedback loops, and governance discipline around that engine.

This reframes what "MT quality" should mean at the procurement stage. A benchmark score at the point of evaluation says little about the three assets that determine long-term production value: proprietary data that enables a system to improve on the organization's own content, integration into the workflows where translation actually happens, and configuration around company-specific terminology, brand voice, and risk tolerance. These are human and organizational assets, not model properties, and they compound over time in ways a static benchmark cannot capture.

The practical implication: A benchmark win recorded today, including the comparative results presented earlier in this document, describes a snapshot of a fast-moving field, not a permanent competitive gap. Locking a procurement decision to the current leaderboard risks optimizing for the one variable most likely to change again within a year.

The more durable capabilities—translation memories and glossaries built from years of human correction, deep integration into existing CAT tools and content platforms, and governance processes that enforce terminology and compliance across every language and department—stay with the enterprise regardless of which model a vendor runs underneath. These are precisely the dimensions this playbook's evaluation framework asks buyers to test in Phases 1 through 5, rather than treating the comparative quality chart as the entire decision.

None of this argues against model quality as a screening criterion; a system that performs poorly on the buyer's own content should still be eliminated early in the process outlined in Phase 2 below. The argument is narrower: model quality functions as a qualifying threshold rather than the deciding factor.

The vendors worth prioritizing for enterprise deployment are the ones that pair strong model performance with the data flywheel, integration depth, and operational tooling that compound in value over time, independent of which specific model generation is running underneath.

The Elements and Implementation of a Robust Framework

Phase 1: Define Enterprise Requirements Before Evaluating

- Content type and use case inventory (legal, marketing, technical, UGC, customer-facing or internal content, etc.)

- Risk profile ("What are the consequences of an [inevitable, but occasional] inaccurate translation? It gets it at *vs.* someone loses an eye")
- Language pair priorities and volume profiles
- Quality tier definitions by use case (with a professional translator quality level linguist post-edit, or gist-only) and clear examples and rules around what constitutes a "good translation"
- Key Enterprise evaluation dimensions:

    - ✔ **General translation quality** — Accuracy and fluency at the segment level
    - ✔ **Terminology adherence** — Consistency with company-specific glossaries, brand terms, and regulated vocabulary
    - ✔ **Tone and voice consistency** — Formal/informal register; brand voice; audience-appropriate language
    - ✔ **Formatting and tag handling** — Preservation of HTML, XML, markdown, placeholders, and layout elements—often catastrophically broken by LLMs without specific requirements tuning
    - ✔ **Full-text consistency** — Coherence across a complete document or product description, not just individual segments
    - ✔ **Primary Goal** — Faster TAT (turnaround time), compliance with guidelines, decrease costs, provide translations for content that would otherwise go untranslated, etc.

- Integration requirements (TMS, API, file formats)
- Data privacy, sovereignty, and security requirements

Phase 2: Construct a Representative Test Environment

- Use your own enterprise content specific to the use case, not vendor-supplied or generic web-scraped test sets
- Recommended scale: 1,000–2,000 segments (10k–20k words) of variable length for automated metrics and 100–250 segments (1k–2.5k words) per language pair for human evaluation
- Ensure test data is genuinely unseen by the systems being evaluated (data leakage audit)
- Establish a double-blind human evaluation panel of a minimum of 3 domain-qualified translators

Phase 3: Use the Right Metrics for Understanding & Selection

- **Automated layer**: COMET + MetricX + CHRF as triangulated signals, not standalone verdicts
- **Human evaluation layer**: Domain-expert double-blind assessment using MQM or similar error taxonomy
- **Human edits measurement layer:** Quantitative measurement of changes made by domain-experts in the form of Levenshtein distance, TER (translation edit rate), or TTE (time to edit)
- **LLM-as-judge layer**: For instruction-following, style guide adherence, and contextual fidelity
- **Business impact layer**: Post-edit productivity, throughput rates, CX impact, conversion/support resolution metrics, A/B tests

Phase 4: Evaluate Adaptability, Not Just Out-of-the-Box Quality

- Run a controlled adaptation trial: provide the vendor with a sample of enterprise data, measure improvement delta over 30/60 days. To efficiently provide this feedback, for decades Translated Srl has utilized specific metrics like Time-To-Edit (TTE) and Errors-Per-Thousand (EPT), which are proven and reliable.
- Assess glossary enforcement (hard constraint vs. instructed suggestion)
- Key question: *How much data, effort, and time does it take to outperform the baseline on our content?*

Phase 5: Assess Production-Readiness, Security, and Governance

- Vendor model update policies: Are you notified? Are you subject to generic LLM model updates and deprecation?
- Data ownership and training rights: Do your corrections data train a shared model?
- Stack ownership: Who owns the training data, architecture, and inference pipeline?
- SLA structure, audit trail, quality accountability
- Data security infrastructure certifications and reliability

Why Human Evaluation Remains Translated's Anchor Standard

- Internal development and enterprise deployments at Translated rely primarily on human assessments, not automated metrics
- The "bottom line" principle for enterprise use: *If it matters to your business success, always include professional human evaluations in your decision-making*

Part 2:
Lara Configuration & Action Playbook

Lara 3: Enterprise Configuration and Setup Guide

Independent analysts have begun formally recognizing that translation management and MT decisions are now board-level, not just localization-team, concerns. Forrester's inaugural Wave for Translation Management Systems (Q3 2025) frames TMSes as enterprise platforms that must orchestrate AI, quality tooling, data security, and human translation options together, not as isolated point tools.

Gartner's research on translation quality at scale confirms that some form of quality estimation and continuous feedback loops are now treated as core enterprise requirements rather than nice-to-haves. For procurement, the practical implication is that vendor evaluation should include how quickly and cheaply an engine adapts to house style and terminology, since this determines total cost of ownership more than per-word pricing alone.

For executive teams, the business case rests on three points:

1. **Faster time to value:** Small data volumes produce measurable quality gains within weeks rather than months, inverting the traditional MT improvement curve.
2. **Compounding returns:** Quality improves with every project because corrections are retained and reused, so cost per word and turnaround improve as volume consolidates with one vendor.
3. **Assured data governance:** The model is developed fully in-house, giving Translated (and by extension, enterprise clients) control over data security and model optimization rather than relying on third-party general-purpose LLM providers.

Recommended Lara Setup

Translated recommends against over-engineering the initial configuration. Lara adapts to translation memory and glossary content automatically, so the practical setup sequence is baseline first, audit second, targeted fixes third.

- **Linguistic Asset Management: Glossaries and Translation Memories**
    - Data hygiene is a prerequisite for model performance, requiring rigorous sanitization of linguistic assets before ingestion.
    - A clean, baseline translation memory (TM) helps Lara internalize specific brand voices, while structured glossaries ensure terminology precision.
    - For most enterprise programs, this disciplined preparation yields high-quality results without extensive manual tuning.
    - Production-ready glossaries must include domain-specific metadata, usage notes, and prohibited terms beyond simple source-target pairs.
    - Explicit "do not translate" (DNT) directives are vital for brand integrity, especially where terms require distinct localized equivalents or must remain untranslated in proprietary strings.
    - Lara's pattern-recognition architecture is sensitive to terminological friction; legacy assets should therefore go through a structured TM Cleanup Pipeline to resolve internal contradictions before onboarding.

- **Governance through Instructions and Quality Settings**
    - Editorial standards are enforced via persistent instructions at the content bucket level, alongside ad-hoc directives for specific job requirements.
    - Active directives are restricted to 10 per request to maintain consistency and prevent quality degradation. Extensive experimentation shows that more than this can result in performance degradation.
    - Translated's data science teams audit model output against existing TM behavior to distill complex brand manuals into high-impact rules. A 40-page brand guide may distill into a handful of persistent instructions such as "convert imperial measurements to metric" or "address the user as *usted* in Spanish".
    - Quality governance relies on the synergy of asset hygiene, glossary precedence, and an exception-based audit workflow to focus organizational efforts.
- **Contextual Fidelity and Asset Prioritization**
    - Lara evaluates full-document context instead of isolated strings, significantly improving pronoun agreement and stylistic coherence across long-form text.
    - The system follows a strict priority hierarchy: Glossary exact (ICE) matches have absolute precedence, followed by TM exact matches.
    - If a perfect match is unavailable, the model dynamically weights all available linguistic assets to generate the most appropriate contextual response.

Human Feedback Operational Loop

Lara's differentiation is built on a carefully refined (over decades) human-machine symbiosis rather than static training. The operational loop runs as follows:

1. **Translation**
2. **Human review**
3. **Correction capture**
4. **Asset update:** corrections feed back into translation memory immediately and, where patterns recur, inform glossary or style-guide refinement.
5. **Adaptation:** Lara continually adjusts its suggestions for subsequent segments within the same project, so editing time decreases as the project progresses.
6. **Quality measurement**: Focused on managing the exceptions.

The combined impact of these measures creates what Translated calls a "data flywheel"—where every delivered project adds professionally corrected data that improves Lara for that client specifically, meaning a client's tenth project measurably outperforms their first.

The table below shows the complete loop from initial translation to sustainable quality gain:

Step	What Happens	Enterprise Impact
1. Translate	Lara assembles context from all seven buckets (source document, TM, glossary, style guide, instructions, external context, and prior corrections) and produces output.	Translation quality reflects the completeness of configured assets from the first segment.
2. Human Review	Linguist works in Matecat alongside Lara. Every edit is observed in real time within the current project, not batched for the next retraining cycle.	Lara updates its suggestions for subsequent segments immediately based on each correction.
3. Correction Capture	Terminology choices, style preferences, and register adjustments propagate forward automatically. A linguist corrects "click" to "tap" in segment 4; Lara applies "tap" from segment 5 onward.	Repetitive editing drops as the project progresses. Linguists focus on ambiguity and nuance, not routine fixes.
4. Asset Update & Adaptation	After the project closes, add confirmed terminology to the glossary. Recurring instructions become Style Guide rules (submitted to Translated for conversion). TM is enriched with corrected segments.	Each project's corrections compound into the next through the data flywheel effect described above.
5. QA Audit	Translated's quality management team audits Lara output against client standards and works by exception: Only the rules Lara is not picking up automatically are added. This keeps the instruction set lean and effective.	CSA Research's 2026 analysis identifies audit-driven, exception-based governance as the differentiator for enterprise-scale global content programs.

Activation Requirements

Integrations. Lara is accessible through TranslationOS (Translated's enterprise delivery platform, where TMs and glossaries are centrally managed), through connectors to major TMS and content platforms, via Model Context Protocol (MCP), and via direct API integration, with functionality varying by access method.

Roles and process. Activation is program manager-led: Enterprise clients work with a Translated program manager to configure the correct linguistic assets per content bucket and to initiate optional pipelines such as TM cleanup or style guide conversion. TM cleanup engagements specifically involve an Account Executive, Account Manager, Program Manager, and Quality Manager during a discovery phase, followed by a scoped statement of work before any cleanup work begins. Program Manager (scoping, asset management); Quality Manager (audit, style guide conversion, exception identification); Data Science (TM Cleanup pipeline execution); Linguists (production translation and correction within Matecat).

Data volumes and cost. Six-step engagement for additional or specialized TM cleanup (if needed): Discovery, Preflight, Run, Health Report, Resolution, Landing. Core rule-based modules (duplicates, language mismatches, placeholder checks, structural errors) are priced at very low prices per translation unit, while advanced AI-based modules (semantic scoring, glossary consistency, tone-of-voice detection) run at higher prices per translation unit. With human-in-the-loop-based resolution, the throughput may run around 100 to 200 units per hour depending on the module.

Adaptation timeline
Small volumes produce significant improvement within weeks. ModernMT's published evaluation research shows adaptive systems outperform static engines on enterprise domain content from the first project; gains compound as TM volume grows. Because Lara adapts fresh per request rather than through periodic retraining, small data volumes can produce significant quality improvements within weeks, with quality continuing to compound as linguistic assets grow. Tone-of-voice detection specifically requires a development phase of approximately 5 to 10 business days to calibrate test prompts before production classification begins. Full program optimization is typically visible within the first several weeks of production.

Concrete Example: Baseline vs. Configured vs. Feedback-Adapted

Using the documented UI login example, the progression illustrates Lara's adaptation model:·

- Baseline (no glossary, mixed TM): The source term for "log in" renders inconsistently, sometimes as "Iniciar sesión" and sometimes as an unrelated variant, because the TM contains conflicting historical patterns for the same term.
- Configured Lara (glossary and clean TM applied): A glossary entry locks the approved term, and a sanitized TM removes the contradictory "Log in" versus "Sign in" segments, so Lara consistently renders the approved terminology from the first segment.
- Feedback-adapted Lara (after human correction): If a linguist corrects something, Lara applies that correction automatically for every subsequent instance in the same project, reducing repetitive edits through the data flywheel effect described above.

This progression is the core commercial argument for procurement: Enterprises are not purchasing a fixed MT engine, but investing in a system whose output improves specifically for their content, with cost and quality both improving as volume and corrections accumulate.

This progression, from inconsistent baseline to rule-enforced configuration to compounding feedback-driven adaptation, is the core basis on which Translated argues its platform reduces both correction time and unit cost as volume with a given enterprise account grows.

Concrete Examples: Baseline, Configured, and Adapted

The table below shows three representative content types, what unconfigured Lara produces, what changes when assets and modes are properly set up, and how output shifts further after the human-feedback loop has run for one or more projects:

Scenario	Baseline Lara (uncofigured)	Configured Lara	After Feedback Adaptation
Software UI (Spanish) Source: "Click to sign in to your Xbox Account."	Lara without glossary: "Haga clic para iniciar sesión en su cuenta de Xbox" ("Xbox Account" translated; "sign in" rendered inconsistently across vendors).	Glossary adds DNT entry for "Xbox Account" and enforces "iniciar sesión" for "sign in." Instruction: "Translate faithfully without creative interpretation." Output: "Haga clic para iniciar sesión en su Xbox Account." Term locked.	Linguist corrects "clic" to "toca" in segment 3 for mobile context. Lara applies "toca" to all remaining segments automatically. Post-edit rate drops by end of document.
Source: "Click to sign in to your Xbox Account." Source: "Discover our new collection."	Lara without Style Guide: "Scopri la nuova collezione" (informal "tu" register; inconsistent across multi-vendor legacy TM).	Style Guide converted to Lara-executable instruction: "Use formal register (Lei) for all marketing content." Output: "Scopra la nuova collezione." Applied automatically to every segment in the bucket without manual intervention.	After the first project enforces "Lei" throughout, TM is enriched with corrected segments. The second project starts with near-zero formality errors. Correction rate continues falling as Lara absorbs the established voice.
Regulated content (French legal) Source: "Do not share your password with others."	TM Cleanup Health Report flags semantic inversion risk: Legacy TM contained "Merci de partager votre mot de passe avec votre équipe" (says the opposite of the source). Without cleanup, Lara learns to reproduce the error.	TM Cleanup AI Semantic Scoring module flags the segment (score near 0). Human linguist corrects and restores. Lara Think enabled: checks output against glossary and context before finalizing. Correct output: "Ne partagez pas votre mot de passe avec d'autres personnes."	Corrected segment added to TM. Lara Think continues to monitor for semantic inversions. Downstream projects on the same content type start with a clean data foundation and no recurrence of the error.

Lara in Action: Playbook for Achieving Continuously Improving Translation Quality

Lara 3 delivered its largest quality jump between two generations by shifting from imitation-based training to reinforcement learning from expert feedback, and enterprises can replicate similar gains on their own content by following a disciplined setup and continuous-improvement workflow.

Three mechanisms sit behind these results: a "learning by doing" training method that lets the model practice, receive structured feedback, and refine itself rather than only copying reference translations; a specialized evaluation model and rigorous human evaluation methodology built on Translated's documented expertise in how professional reviewers actually judge translations; and purpose-built infrastructure, that supports both training and continuous customer adaptation.

Crucially, Translated maintains a clear separation between model training (the LLM itself) and customer adaptation (how Lara learns a specific enterprise's terminology, TM, and style through use), which is why quality can keep improving between projects without retraining the base model.

1. How Lara Achieved the Results

Learning by doing

Earlier Lara versions, like most MT systems, trained by imitation—shown a single reference translation per sentence and rewarded for matching it, even though any sentence can be correctly translated in more than one way. **Lara 3 replaces this with a three-stage practice loop built into final-phase training: The model explores alternative translations, an automatic judge reviews each for what works and why, and Lara refines its choices based on that feedback, with the loop repeated millions of times across the training corpus.** Translated frames this as part of the shift in AI, where models learn from generated experience rather than only historical data, mirroring how human translators develop judgment through practice and reviewer feedback

Specialized evaluation model

The automatic judgment that scores Lara's practice attempts was built on Translated's documented expertise in how professional reviewers actually evaluate translations, including their correction patterns and terminology standards, rather than a generic scoring function. This is a meaningful distinction for teams evaluating vendor claims: A judge trained on real reviewer behavior produces reward signals that correlate with what a human buyer would actually accept, rather than a proxy metric optimized for research convenience.

Purpose-built infrastructure

Translated's specialized hardware investments provide the infrastructure basis for Lara's simultaneous quality, cost, and speed advantages. Unlike general-purpose LLMs, which lack a reliable mechanism for enforcing consistent terminology across thousands of segments in a document, Lara's architecture uses enterprise-grade retrieval and attention mechanisms trained exclusively on high-quality linguistic data rather than general web content. This infrastructure allows Lara to avoid making quality, cost, or speed trade-offs that generic LLM use would require.

Model training vs. customer adaptation

Model training happens once, centrally, and produces the base Lara model; customer adaptation happens continuously and separately, as Lara retrieves and applies each client's own translation memory, glossary, style guide, and real-time corrections at the moment of translation. This is why small data volumes can produce measurable quality gains within weeks rather than the months typical of generic approaches.

2. Configure Lara for Your Content

Establish the baseline. Translated recommends against over-engineering initial setup: Start with existing translation memories and glossaries, since Lara adapts to them automatically and this alone produces strong results for most programs.

Prepare and clean linguistic assets. Translation memories are subject to progressive structural degradation. Over years of shifting vendor partnerships, evolving brand standards, and decentralized editorial interventions, these repositories often accumulate contradictory patterns that compromise production output. Because Lara internalizes knowledge directly from available data, this terminological noise forces the model to reproduce errors, while linguists lose substantial throughput reconciling conflicting matches. Left unaddressed, these inconsistencies act as a silent tax on every future project, inflating post-editing effort and driving up the total cost of ownership.

The TM Cleanup Pipeline applies a disciplined, data-driven optimization layer to enterprise assets, utilizing the same rigorous sanitization protocols that Translated employs for its own model training. By auditing and refining legacy data, the pipeline establishes a high-quality baseline that improves initial output, compounds cost efficiencies, and ensures that linguistic records function as reliable strategic assets rather than repositories of historical friction.

Manual alternatives, such as random sampling or bulk deletions, are either too slow to scale or risk discarding valuable proprietary content. In contrast, this pipeline identifies specific points of failure—including semantic inversions and placeholder errors—to target remediation where it yields the highest operational impact. For global organizations, the strategic advantages of this data-centric approach include:

- Superior translation accuracy and terminological consistency across the global language portfolio
- Reduced unit costs through increased match confidence and minimized repetitive post-editing effort
- AI-ready training data that enables Lara to deliver specialized, domain-aware output from day one
- Enhanced linguistic focus, allowing professional reviewers to work within a reliable and governed data environment

The optional TM Cleanup Pipeline runs automated checks (duplicates, language mismatches, placeholder and tag integrity), AI filtering (cross-entropy scoring for low-quality segments), and human linguist review for ambiguous cases before the cleaned TM feeds Lara's adaptation layer. The full engagement runs through the six-step TM Cleanup Pipeline (described previously) with a defined owner and exit gate at each stage.

Build glossaries. Entries should include source and target terms plus domain, subdomain, usage notes, and do-not-translate (DNT) flags; for example, "account" resolves to "cuenta" in finance content, "cuenta de usuario" in gaming content, and as illustrated in the previous DNT example. Glossary exact (ICE) matches take absolute precedence over every other asset Lara consults.

Convert style guides into instructions. A 25-page brand manual is typically distilled into a handful of persistent, Lara-executable instructions, such as "convert imperial measurements to metric" or "address the user as *usted* in Spanish," because Lara already absorbs many stylistic patterns automatically from the TM. Style guide rules and ad-hoc instructions share one field capped at 10 total directives per request; exceeding this makes output stilted and rules inconsistently applied.

Select Standard Lara, Lara Think, or Lara Prosa. Standard Lara suits high-volume, well-configured content like UI strings. Lara Think runs multi-step reasoning against glossaries and style guides, catching most major terminology and style issues before human review, at the cost of longer processing time, making it appropriate for regulated or technical content. Lara Prosa applies an explicit editorial style guide to preserve tone, rhythm, and voice, suited to marketing and literary material where brand identity is paramount.

Configure context and integrations. Lara processes seven context buckets: source document, translation memory, glossary, style guide, instructions, external context, and real-time corrections. Deployment runs through TranslationOS, connectors to major CAT tools (Trados, memoQ, Phrase, Crowdin, Wordfast), direct API/SDK access, or an MCP server linking to AI assistants, with functionality varying by access method.

Test with representative content. Run a pilot on your own content rather than vendor-supplied samples, since a small, generic test set produces statistically fragile results; a credible enterprise evaluation uses a meaningful sample size and is reviewed by at least three domain-qualified translators.

3. Activate Continuous Improvement

When the workflow is appropriate. The human-feedback loop is most valuable for recurring content types with sustained volume, UI strings, product documentation, support content, and marketing copy, where corrections compound across many future segments and projects rather than one-off translations.

Activation prerequisites. Activation is program-manager-led: the enterprise client works with a Translated program manager to configure the appropriate linguistic assets for each content bucket and to initiate optional pipelines such as TM cleanup or style guide conversion. Roles typically involved include the Program Manager (scoping and asset management), Quality Manager (audit and exception identification), Data Science (pipeline execution), and Linguists (production translation and correction).

Human review in Matecat. Linguists review and correct Lara's output inside Matecat, Translated's live CAT tool, with every edit observed in real time.

Real-time project adaptation. A correction made in segment 12 changes what Lara proposes from segment 13 onward within that same project, so linguists never repeat the same fix twice.

Validation and asset updates. After a project is closed, confirmed terminology gets added to the glossary, recurring instructions become style guide rules, and corrected segments enrich the TM, so gains persist into future projects rather than resetting each time.

Cross-project quality improvement. Translated's quality management team periodically audits output against client standards, working by exception to add only the rules Lara is not already absorbing automatically, keeping the instruction set lean.

Roles, governance, and measurement. Ongoing governance combines asset hygiene, glossary precedence rules, and exception-based audits, an approach CSA Research's 2026 analysis identifies as the differentiator for enterprise-scale global content programs.

Role	Responsibility
Program Manager	Scoping, asset management, TM cleanup coordination
Quality Manager	Audit, style guide conversion, exception identification
Data science	TM Cleanup pipeline execution
Linguists	Production translation and correction within Matecat

3. Activate Continuous Improvement

When the workflow is appropriate. The human-feedback loop is most valuable for recurring content types with sustained volume, UI strings, product documentation, support content, and marketing copy, where corrections compound across many future segments and projects rather than one-off translations.

Activation prerequisites. Activation is program-manager-led: the enterprise client works with a Translated program manager to configure the appropriate linguistic assets for each content bucket and to initiate optional pipelines such as TM cleanup or style guide conversion. Roles typically involved include the Program Manager (scoping and asset management), Quality Manager (audit and exception identification), Data Science (pipeline execution), and Linguists (production translation and correction).

Human review in Matecat. Linguists review and correct Lara's output inside Matecat, Translated's live CAT tool, with every edit observed in real time.

4. Configuration Paths by Use Case and Content Type

Use Case / Content Type	Context Priority	Mode / Settings	Example Instruction
Software / UI	Glossaries (DNT entries critical), TMs	Standard Lara + strict instructions	*"Translate faithfully without creative interpretation." Glossary: "Xbox Account" = DNT; "sign in" = "iniciar sesión" (ES).*
User-Generated Content	User-Generated Content	Standard Lara + profanity moderation	*Profanity Detection enabled as a standalone moderation layer. Instructions: "Regional variant: es-MX."*
Technical Documentation	TMs, glossaries (domain + subdomain), External Context	Lara Think for regulated/certified content	*"Technical audience. Preserve all variable placeholders." Lara Think flags ~80% of major terminology mismatches before output.*
Marketing / Brand	Style Guides (formality, register), TMs	Lara Prosa for literary/premium campaigns	*"Use a formal register. Address user as 'usted' in Spanish." Lara Prosa enforces editorial voice and rhythm across brand copy.*
Regulated / Legal	Glossaries (approved terms), TMs (clean), Style Guides	Lara Think required; HITL review mandatory	*"Translate literally without interpretation." Lara Think checks output against glossaries and context before finalizing, catching semantic inversions.*

A documented example illustrates the stakes for regulated content: A legacy French legal TM contained "Merci de partager votre mot de passe avec votre équipe," which says the opposite of the intended source, "Do not share your password with others." The TM Cleanup semantic scoring module flagged the segment near a zero similarity score, a human linguist corrected it, and Lara continued monitoring for recurrence across downstream projects.

5. Results Measurement Framework

CSA Research's language risk framework argues that **the appropriate quality bar is not the highest achievable linguistic score, but the level sufficient to avoid harm to customers, brand, or legal standing,** which is why enterprise measurement should be tied to business fitness rather than a single number. Translated itself relies primarily on human evaluation rather than automated metrics for its own quality decisions. Forrester's localization research finds organizations fail more often from workflow and adoption misfit than from linguistic imperfection. Recommended metrics include:

- **Human approval rate**: the percentage of sentences where a majority of independent professional reviewers approve the translation as suitable for professional use, the same double-blind methodology used in Lara 3's comparative evaluation.
- **Time to edit (TTE)**: the actual editing effort a linguist needs to bring MT output to publish quality, proposed by Translated as more production-relevant than static reference-based scores because it ties directly to cost and throughput.
- **Terminology adherence**: consistency with company-specific glossaries, brand terms, and regulated vocabulary, measurable through glossary hit rate and manual audit of DNT compliance.
- **Style compliance**: adherence to formality register, tone, and brand voice rules, assessed through the same audit-driven, exception-based process used to build style guide instructions.
- **Error recurrence**: whether a corrected error reappears in later segments or projects, tracked via Errors Per Thousand (EPT) and reviewed during periodic quality audits.
- **Improvement between projects**: the compounding delta in TTE, EPT, or approval rate from a client's first project to subsequent ones, the core evidence for the data flywheel claim.

6. Launch Checklist

Before going live, confirm TM has completed cleanup or a documented waiver, glossary includes DNT and domain-variant entries, style guide has been converted into fewer than 10 persistent instructions, the appropriate mode (Standard, Think, or Prosa) is selected per content bucket, integration and roles are assigned, and a representative pilot has been run and reviewed.

Summary Operational Checklist for Localization Managers

Before activation

- Confirm content bucket and priority use case (UI, documentation, marketing, UGC, legal)
- Pull existing TM and glossary; flag for TM Cleanup Pipeline if legacy data has multi-vendor history or known contradictions
- Submit Data Science Request Form (client name, problem statement, modules, scope, TMX files, deadline) if cleanup is needed; allow buffer time, no same-day requests

Asset setup

- Glossary: Add source/target terms, domain and subdomain, usage notes, DNT entries
- Style guide: Work with program manager to convert into a short list of persistent instructions (target under 10 total directives combined with ad-hoc instructions)
- Select mode: Standard Lara (general), Lara Think (regulated/technical), Lara Prosa (brand/marketing voice)

Pilot and validation

- Test on 100–250 segments per language pair for human review; 1,000–2,000 for automated metrics, using your own content, not vendor samples
- Use a minimum 3-translator double-blind panel for quality sign-off
- Review TM Cleanup Health Report (flagging rate by module, similarity score distribution) before approving further resolution work

Production and continuous improvement

- Confirm linguists are reviewing inside Matecat with real-time correction capture enabled
- Schedule periodic exception-based audits with Quality Manager; add only missing rules, not full rewrites
- After each project, push confirmed corrections into glossary/TM/style guide for reuse in the next project
- Track TTE, human approval rate, terminology adherence, and project-over-project improvement as standing KPIs

Governance

- Confirm role-based asset control: who can edit glossaries/TM vs. who can only use them
- Confirm model-update and data-ownership terms, especially for regulated content lines
- Reconfirm cleaned-TM delivery method (FTP, TMS, or TranslationOS) and that only touched segments were altered

The Translated/Lara Value Proposition

How Lara Is Purpose-Built for the Enterprise

Positioned as the successor to ModernMT, a pioneering adaptive machine-translation technology launched in 2016, Lara represents the culmination of a decade of innovation. Translated's original adaptive engine was the first to provide real-time commercial capabilities, later enhanced with transformer-based neural architectures in 2017. This technical lineage has been consistently recognized by leading industry analysts: Translated has secured recognition as a technology leader by Gartner, IDC MarketScape, and CSA Research. This trajectory underscores a deep-rooted expertise in developing high-performance, domain-specific AI solutions that deliver measurable advantages to global organizations and professional linguists.

ModernMT's technical documentation makes a point that carries through into Lara's design: Most static MT systems require significant upfront investment before they adapt to a customer's domain and data, while an adaptive architecture has a natural advantage because it is built to adapt and adjust from the start. Lara carries that architecture into a large language model built specifically for translation rather than repurposed from a general one.

Lara, developed by Translated, is the largest large language model (LLM) built exclusively for translation, trained on more than 10 million GPU hours of proprietary linguistic data and supporting over 206 languages.

General-purpose LLMs are not designed to ingest linguistic assets like translation memories and glossaries, and they lack a reliable mechanism for enforcing the same terminology and phrasing across thousands of segments in a document or project. Lara is trained exclusively on high-quality linguistic data that lets it prioritize the most relevant patterns from a client's own content.

For enterprise buyers evaluating machine translation and localization technology, Lara's core differentiator is not raw model size but three connected capabilities:

1) fast startup, low-friction setup,

2) a disciplined and rigorous translation memory (TM) cleanup process instead of assuming that data is already usable, and

3) a structured human feedback loop that compounds quality improvements over time. Independent validation from CSA Research, a leading globalization and localization market research firm, describes Lara as a move toward "responsible" machine translation that learns from both good translations and mistakes.

For enterprise procurement stakeholders, the decision-making framework for MT typically centers on three operational pillars: deployment velocity, total cost of ownership, and the long-term quality trajectory. The integration of rapid, low-friction setup, rigorous data validation, and a structured human feedback loop that compounds quality gains throughout the program lifecycle, rather than relying on static retraining intervals, positions Lara as a strategic infrastructure choice that addresses all three requirements simultaneously.

Straightforward Setup: A Lower-Risk Path to Production

- Enterprise stakeholders typically face significant implementation barriers, including integration complexity, data sanitization, and heavy resource lifts to reach production-ready quality, which Lara addresses through a rapid-deployment architecture and strong out-of-the-box performance.
- The platform automatically internalizes organizational stylistic nuances and terminology by autonomously retrieving patterns from a client's existing translation memories and glossaries, eliminating the overhead of manual rules.
- Translated recommends an audit-driven, exception-based onboarding approach: establish a baseline using current assets, benchmark the output against standards, and target only the identified gaps, which accelerates production timelines and minimizes implementation costs by preventing over-engineered style guides.
- Lara integrates seamlessly into existing infrastructure via TranslationOS for managed human-in-the-loop workflows, native connectors for popular CMS and TMS platforms, a comprehensive API and SDK suite, and an MCP server that links to AI assistants like Claude and ChatGPT.

Translation Memory Cleanup: Turning Legacy Data into an Asset

Legacy translation assets are seldom prepared for immediate integration into an LLM ecosystem. The cumulative effect of multi-vendor editing cycles often results in repositories characterized by errors and obsolete records; since Lara functions through sophisticated pattern-recognition, these internal contradictions can actively compromise the precision of generated output.

To mitigate this, Translated employs a specialized data sanitization pipeline that processes these linguistic assets through three distinct phases before model ingestion:

Stage	What it does
Automated checks	Semantic similarity, language mismatches, terminology conflicts, placeholder and tag integrity, punctuation, and duplicate removal.
AI filtering	Cross-entropy scoring and reasoning supervision identify and flag low-quality or unreliable segments.
Data science	Professional linguists resolve the ambiguous, low-confidence cases automation cannot settle on its own.

Lara's adaptation layer ingests this sanitized and AI-ready memory directly, converting decades of previous linguistic choices into a robust functional asset on day one. By employing this disciplined cleanup process, organizations transform their legacy translation history into a primary competitive differentiator, ensuring that historical records serve as a foundation for accuracy rather than a repository of persistent errors.

Translated also offers additional data optimization services that provide more focused deep cleaning of translation memory to ensure minimal impact on downstream AI system use. These services require thorough problem identification and a pre-agreed data science-based plus human review resolution process that ensures that data is a long-term AI training asset.

The Human Feedback Operational Loop: Compounding Quality Through Continuous Learning

For executive stakeholders concerned with long-term vendor lock-in and diminishing returns, the durability of Lara's improvement mechanism is a central argument. Lara operates alongside human linguists in real time within Translated's Matecat editing environment, observing every correction a translator makes and applying that learning to subsequent segments within the same project immediately, not at the next training cycle. CSA Research's independent evaluation validated this as a genuine architectural advance, noting that Lara "learns from both good translations and mistakes" and that Translated's retention of what it calls "arbitration data"—the intermediate reasoning and feedback behind translation decisions—provides insight that conventional TM repositories cannot capture. The same report states that this creates "a long-term advantage" over MT systems that produce only static output.

As a project unfolds, terminology consistency and stylistic nuances are autonomously internalized, ensuring linguists never repeat the same manual correction. This dynamic reinforcement creates a compounding operational advantage, where translation quality improves cumulatively through continuous learning.

For procurement functions, this carries a significant commercial weight: Translation quality is an incremental and compounding advantage rather than a reset with every engagement. Major global organizations—including **Amazon, Airbnb, Asana, and Notion**—centralize their localization volume within this ecosystem because increased usage yields statistically superior performance, extending beyond simple per-word savings. Asana realized a 70% automation rate and a 30% reduction in manual oversight by pivoting to this AI-driven architecture, while Airbnb successfully scaled into 31 new language markets. This transition establishes Lara as an appreciating enterprise asset rather than a transactional utility, a value proposition designed for finance stakeholders and executive decision-makers assessing strategic, multi-year partnerships.

For IT and procurement stakeholders, assessing this technology necessitates a departure from standard software acquisition queries. The relevant inquiry extends beyond immediate benchmark results to the deployed operational lifecycle: the transformation of legacy translation memories during setup, the level of team effort required before launch, and the rigor of the ongoing human correction loop. Organizations must evaluate whether a vendor provides verifiable evidence of compounding quality gains or merely maintains a static performance baseline as volume increases. The strategic imperative has shifted; the decision now centers on identifying which vendor's data architecture and refinement processes will yield the most durable long-term returns.

Conclusion: From Evaluation Competence to Enterprise Confidence

The value proposition for executive stakeholders centers on a unified triad of capabilities: a foundation model trained on two decades of specialized production data, a disciplined methodology for sanitizing legacy assets during onboarding, and a structured reinforcement loop that yields cumulative quality improvements. This integrated architecture allows global organizations to validate performance through a controlled pilot before scaling their enterprise-wide commitment.

The Competitive Advantage of Rigorous MT Evaluation

- The enterprise that evaluates machine translation correctly gains a durable competitive advantage in global markets. Moving beyond outdated string-matching metrics and rigid error-counting toward multidimensional evaluation frameworks—such as COMET-DA paired with human expert validation—gives leadership true visibility into translation accuracy and precision. By accurately measuring "shippability," brand alignment, and market readiness, organizations eliminate the silent failures of unguided AI, launching faster in new regions with absolute trust.

Transforming MT into Strategic Infrastructure

- The right evaluation framework transforms machine translation from a tactical cost-center experiment into a high-value strategic investment. By abandoning transactional per-word metrics in favor of business outcome indicators, localization shifts from a back-office expense to a growth engine. Deployed through unified orchestration platforms like TranslationOS, enterprise AI scales content volume seamlessly across all digital touchpoints while driving down long-term operational costs.

Translated Srl: A Long-Term Partner in Global Growth

- The proven lineage of ModernMT, alongside consistent recognition from Gartner, IDC, and CSA Research, provides the objective validation necessary for executive and procurement stakeholders to greenlight enterprise-scale investments. Beyond commodity service, Translated positions itself as a long-term strategic partner by maintaining full end-to-end control over its proprietary AI, Lara. This vertical integration shields organizations from the volatility of model drift, pricing fluctuations, and the governance vulnerabilities associated with general-purpose public APIs. Within a closed-loop environment, professional human review acts as a continuous training signal, internalizing your specific brand voice and terminology. For decision-makers, this ensures a context-aware infrastructure that improves cumulatively, transforming translation into a durable enterprise asset.

Step into
AI - driven localization

Our team is ready to find a solution
for your translation needs

Hello, I'm John.
How can I help you?

John Tinsley - VP of AI Solutions
enterprise@translated.com

translated.