# Unlocking Enterprise Translation Hyper-Automation with LLMs

AMTA Keynote

Alon Lavie
VP of AI Research
October 2024

Phrase

# Acknowledgements

Joint work with the **Phrase AI Research team**: Craig Stewart, Aleš Tamchyna, Silja Hildebrand, Miroslav Štola, Ondřej Glembek, Alan Eckhardt, Sergio Penkale, Val Ulitin, Daniel Martín-Albo Simón, Zuzana Šimečková and Rachel Grewcock

# The Emerging New Enterprise Localization Landscape

Generative AI is driving **massive growth in content creation** and generation ⇒ massive **growth in demand for content localization** but under fixed budgets

Neural and LLM-based Machine Translation **has become very good**

Multilingual Enterprises and Translation Technology and Service Providers have been **racing to adapt the way they do translation at scale**.

The MT+PE paradigm of “First MT then Edit/Review Everything” is **broken and does not scale**

# The Emerging New Enterprise Localization Landscape

**The Challenge**: how can enterprises **adopt automated translation at scale while guarding against the quality risks**, and balance quality/cost/speed with new automated quality evaluation processes?

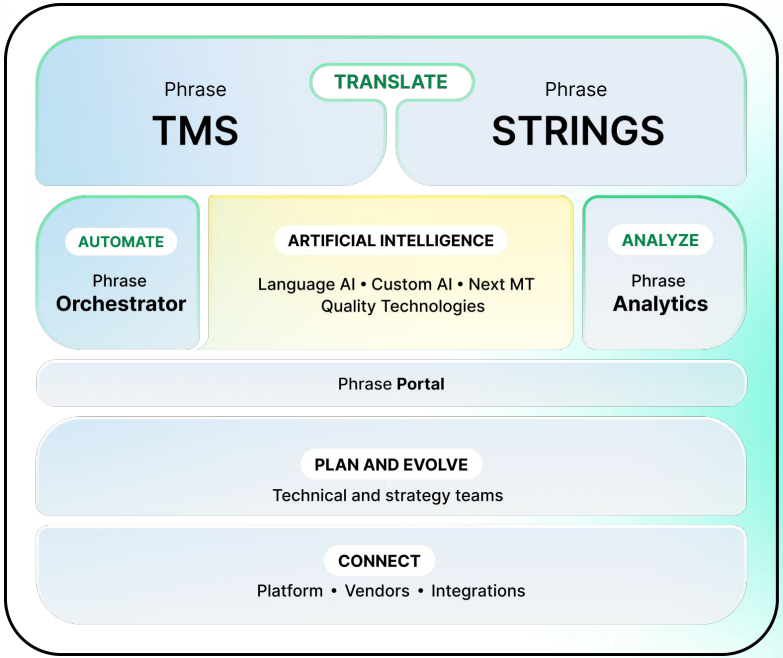
The emerging secret key to unlocking large scale translation automation:

**Automated Translation Quality Technology (TQT) at scale**

**Can GenAI and LLMs offer us exciting new solutions to this challenge?**

# The Phrase Translation/Localization Platform

- **Broad cloud-based Translation/Localization Platform** anchored on solid foundations
- Focus on **Automation workflows**
- **Platform only** - for both Enterprises and LSPs
- **Strong Technology Aggregator** positioning:
  - Support large range of component solutions wherever possible
  - 30+ MT API integrations + Phrase's internally-developed NextMT and NextGenMT as a managed service
- **Customers are not siloed** to just one solution
- Especially valuable during this period of **rapid transformation driven by GenAI**



# Phrase's AI Research Mission

As the AI landscape shifts towards an ecosystem of LLMs, we will evolve our product to provide a platform for localization that:

Provides **best-in-class LLM-based features** that solve genuine customer problems

**Drives automation** to the upper ceiling and **optimizes** the use of human translators.

Our AI hub provides **a framework for understanding, interpreting and acting on translation quality** issues that seeks to **minimize the risk** of optimizing using LLM-based solutions

# An Ecosystem of Quality Technology

## The backbone of a hyper-automated localization workflow

**Automated Scoring**

Dynamically evaluate the quality of your output

Identify and classify issues

# An Ecosystem of Quality Technology

## The backbone of a hyper-automated localization workflow

### Automated Scoring

Dynamically evaluate the quality of your output

Identify and classify issues

### Automated Routing

Use quality signals to ring-fence and route content to the most appropriate and valuable workflow steps

# An Ecosystem of Quality Technology

## The backbone of a hyper-automated localization workflow

### Automated Scoring

Dynamically evaluate the quality of your output

Identify and classify issues

### Automated Routing

Use quality signals to ring-fence and route content to the most appropriate and valuable workflow steps

### Automated Fixing

Correct errors, and enforce consistency in style and terminology

# Phrase Quality Technology

Game-changing translation technology

## QPS

Quality Performance Score: a lightweight, scalable AI scoring mechanism for capturing linguistic quality.

QPS is our primary gating mechanism for routing content and providing a holistic view on global quality.

## Auto LQA

A complement or alternative to human Linguistic Quality Assurance.

LLM-powered, MQM-based assessment that exposes errors and the nature of these errors at a granular level.

## Auto Review

A tool for automated correction, powered by Generative AI, and focused on style and terminology consistency.

Maximize the fitness-for-purpose of translation at scale.

(Coming December)

# What is Translation Quality Estimation?

Given **a source language text** and **a translation**; **automatically** provide an assessment of the **quality of the translation**

## Word-level

**For each word in the translation decide which ones are ‘bad’**

## Quality Score Prediction

**Provide a number that tells how how ‘good’ the translation is**

# What is Translation Quality Estimation?

Given **a source language text** and **a translation**; **automatically** provide an assessment of the **quality of the translation**

## Word-level

**For each word in the translation decide which ones are ‘bad’**

**S: This is an example**

**T: Questo è un taco**

## Quality Score Prediction

**Provide a number that tells how how ‘good’ the translation is**

**0/100**

# What is Translation Quality Estimation?

Given **a source language text** and **a translation**; **automatically** provide an assessment of the **quality of the translation**

## Word-level

**For each word in the translation decide which ones are ‘bad’**

**S: This is an example**

**T: Questo è un taco**

## Quality Score Prediction

**Provide a number that tells how how ‘good’ the translation is**

**0/100**

**We now have AI technology that can do this automatically at translation time reliably!!**

# Applications of QE in Translation Workflows

**Intelligent Routing** of translated content to human editing (MTPE) based on dynamic on-the-fly MT quality analysis, **at both the job and segment level**

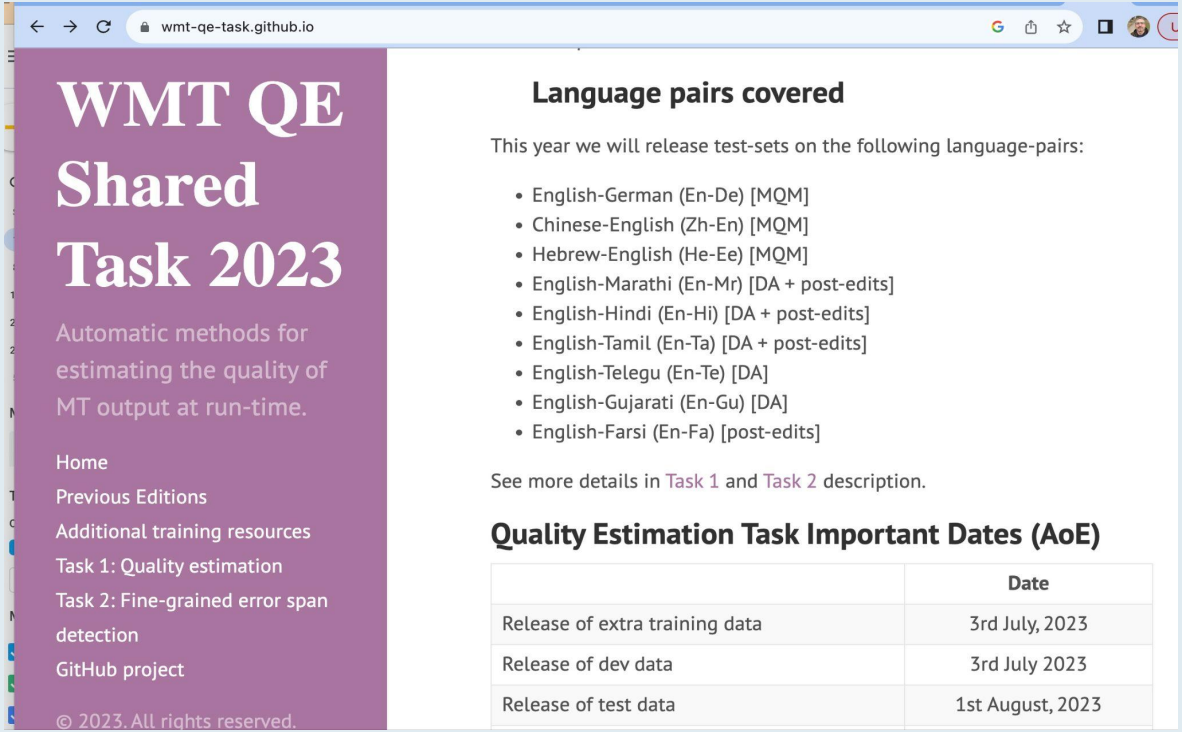
**QE-based MT Benchmarking** allows immediate contrastive quality comparison of different MT systems on any given source content at translation time

**Dashboards** providing full translation quality transparency for all content being translated through a translation platform (both MT and HT).

**QE-based assessment and monitoring of human translators and editors and their work**.

And more...

# The Standard for Assessing QE Accuracy is Based on Multidimensional Quality Metrics (MQM)



# MQM 2.0 (Arle Lommel and the MQM Council)

**MQM (Lommel *et al.*, 2014)**

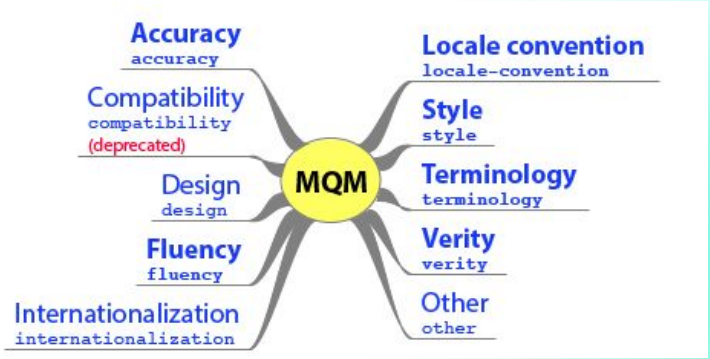
A translation error annotation and scoring framework

- A standardized Error Typology
- Error severity classification
- A quality scoring schema

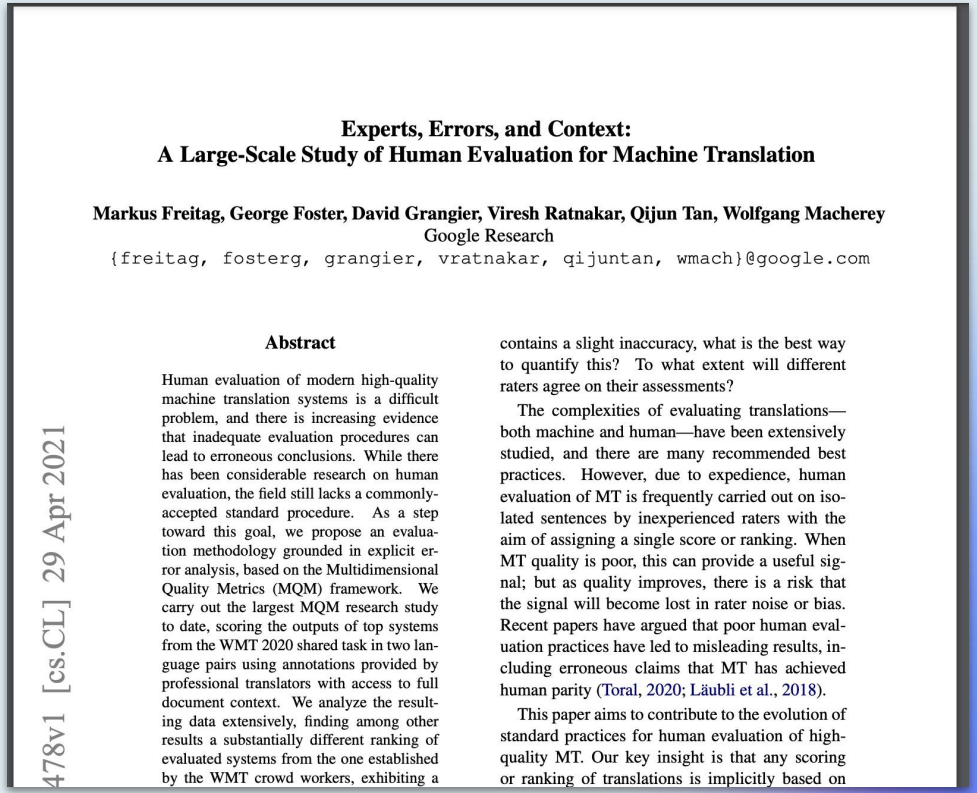
Focus on objective quality assessment in a short amount of time
Allows refinement and customization

Increasing adoption in the translation industry for Human LQA

- MQM 2.0 (2022):  updated version of the framework
- Forms the basis of the new ISO-5060



# Google's 2021 MQM Data Study Paper



**Experts, Errors, and Context:**
**A Large-Scale Study of Human Evaluation for Machine Translation**

**Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, Wolfgang Macherey**
Google Research
{freitag, fosterg, grangier, vratnakar, qijuntan, wmach}@google.com



**Abstract**

Human evaluation of modern high-quality machine translation systems is a difficult problem, and there is increasing evidence that inadequate evaluation procedures can lead to erroneous conclusions. While there has been considerable research on human evaluation, the field still lacks a commonly-accepted standard procedure. As a step toward this goal, we propose an evaluation methodology grounded in explicit error analysis, based on the Multidimensional Quality Metrics (MQM) framework. We carry out the largest MQM research study to date, scoring the outputs of top systems from the WMT 2020 shared task in two language pairs using annotations provided by professional translators with access to full document context. We analyze the resulting data extensively, finding among other results a substantially different ranking of evaluated systems from the one established by the WMT crowd workers, exhibiting a

contains a slight inaccuracy, what is the best way to quantify this? To what extent will different raters agree on their assessments?

The complexities of evaluating translations—both machine and human—have been extensively studied, and there are many recommended best practices. However, due to expedience, human evaluation of MT is frequently carried out on isolated sentences by inexperienced raters with the aim of assigning a single score or ranking. When MT quality is poor, this can provide a useful signal; but as quality improves, there is a risk that the signal will become lost in rater noise or bias. Recent papers have argued that poor human evaluation practices have led to misleading results, including erroneous claims that MT has achieved human parity (Toral, 2020; Läubli et al., 2018).

This paper aims to contribute to the evolution of standard practices for human evaluation of high-quality MT. Our key insight is that any scoring or ranking of translations is implicitly based on

# QE Systems as MT Metrics are now State-of-the-Art

## Results of WMT23 Metrics Shared Task: Metrics might be Guilty but References are not Innocent

**Markus Freitag**[1], **Nitika Mathur**[2], **Chi-kiu Lo 羅致翹**[3], **Eleftherios Avramidis**[4], **Ricardo Rei**[5,6,7], **Brian Thompson**[8], **Tom Kocmi**[9], **Frédéric Blain**[10], **Daniel Deutsch**[1], **Craig Stewart**[11], **Chrysoula Zerva**[7,12], **Sheila Castilho**[13], **Alon Lavie**[11], **George Foster**[1]
[1]Google Research [2]Oracle Digital Assistant [3]National Research Council Canada
[4]German Research Center for Artificial Intelligence (DFKI) [5]Unbabel [6]INESC-ID
[7]Instituto Superior Técnico [8]AWS AI Labs [9]Microsoft [10]Tilburg University [11]Phrase
[12]Instituto de Telecomunicações [13]Dublin City University
wmt-metrics@googlegroups.com

**Abstract**

This paper presents the results of the WMT23 Metrics Shared Task. Participants submitting automatic MT evaluation metrics were asked to score the outputs of the translation systems competing in the WMT23 News Translation Task. All metrics were evaluated on how well they correlate with human ratings at the system and segment level. Similar to last year, we acquired our own human ratings based on expert-based human evaluation via Multidimensional Quality Metrics (MQM). Following last year's success, we also included a challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors. Furthermore, we improved our meta-evaluation procedure by considering fewer tasks and calculating a global score by weighted averaging across the various tasks.

We present an extensive analysis on how well metrics perform on three language pairs: Chinese→English, Hebrew→English on the sentence-level and English→German on the paragraph-level. The results strongly confirm the results reported last year, that neural-based

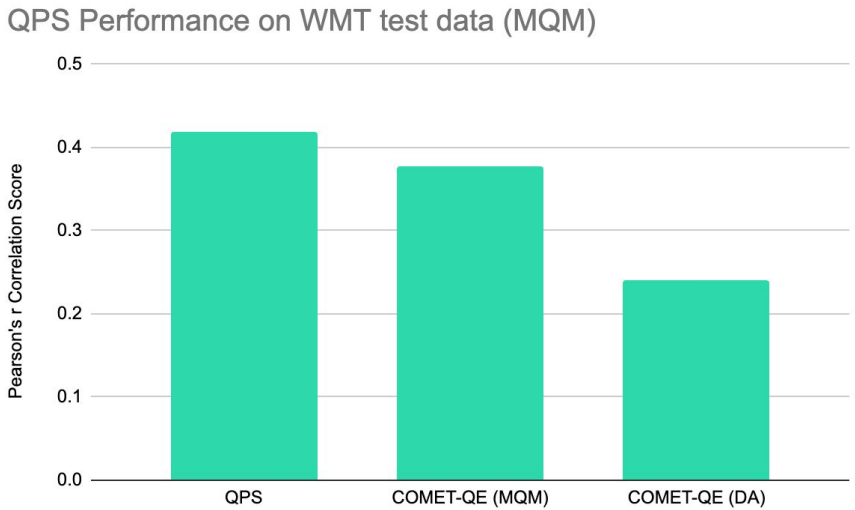
| Metric | | avg corr |
|---|---|---|
| XCOMET-Ensemble | **1** | **0.825** |
| XCOMET-QE-Ensemble* | 2 | 0.808 |
| MetricX-23 | 2 | 0.808 |
| GEMBA-MQM* | 2 | 0.802 |
| MetricX-23-QE* | 2 | 0.800 |
| mbr-metricx-qe* | 3 | 0.788 |
| MaTESe | 3 | 0.782 |
| CometKiwi* | 3 | 0.782 |
| COMET | 3 | 0.779 |
| BLEURT-20 | 3 | 0.776 |
| KG-BERTScore* | 3 | 0.774 |
| sescoreX | 3 | 0.772 |
| cometoid22-wmt22* | 4 | 0.772 |
| docWMT22CometDA | 4 | 0.768 |
| docWMT22CometKiwiDA* | 4 | 0.767 |
| Calibri-COMET22 | 4 | 0.767 |
| Calibri-COMET22-QE* | 4 | 0.755 |
| YiSi-1 | 4 | 0.754 |
| MS-COMET-QE-22* | 5 | 0.744 |
| prismRef | 5 | 0.744 |
| mre-score-labse-regular | 5 | 0.743 |
| BERTscore | 5 | 0.742 |
| XLsim | 6 | 0.719 |
| f200spBLEU | 7 | 0.704 |
| MEE4 | 7 | 0.704 |
| tokengram_F | 7 | 0.703 |
| embed_llama | 7 | 0.701 |
| BLEU | 7 | 0.696 |
| chrF | 7 | 0.694 |
| eBLEU | 7 | 0.692 |
| Random-sysname* | 8 | 0.529 |
| prismSrc* | 9 | 0.455 |

# Phrase's QE Technology: QPS

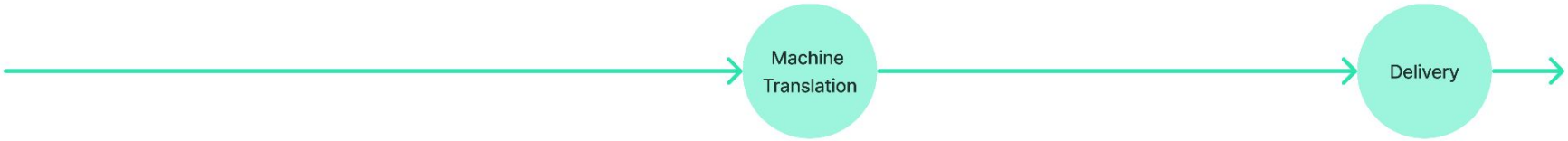
- **Quality Performance Score (QPS)** is Phrase's Quality Estimation (QE) model and system
- **Neural technology** - built on top of Meta's pre-trained multilingual XLM-Roberta model
- **Trained on MQM-annotated human LQA data** (public and proprietary)
- **Calibrated** to generate normalized MQM scores [0-100]
- **Optimized for accuracy** on our primary customer content-types and use-cases
- **MQM is the industry standard** for both human evaluation (the new ISO-5060) and is increasingly used as a "gold standard" in automated translation evaluation.

# Evaluation of QPS

Evaluating automated quality metrics involves relying on human output as a gold standard. This is challenging because even human annotators don't always correlate well amongst themselves. On public test data from Google, human annotators correlate on average at 0.47 Pearson.



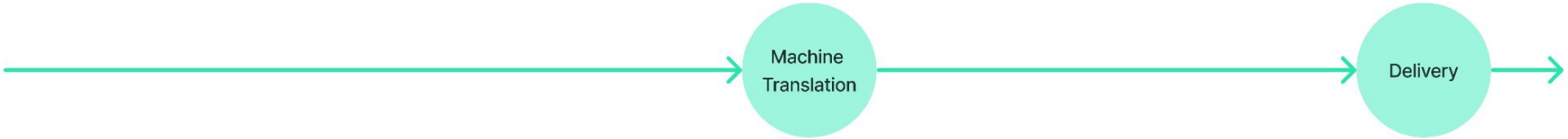
# Traditional MT Automated Workflows




Many high-volume **customers rely on pure MT solutions** to accommodate high volumes at low cost.

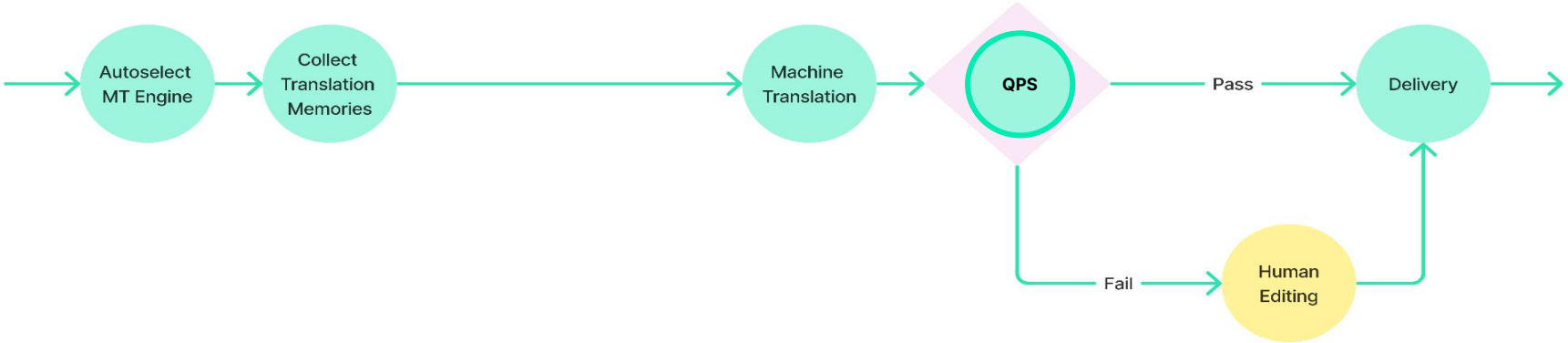
For these customers, being blind to quality and treating all content in the same way presents an unacceptable risk.

# Automated workflows with QPS-based Job-level Routing



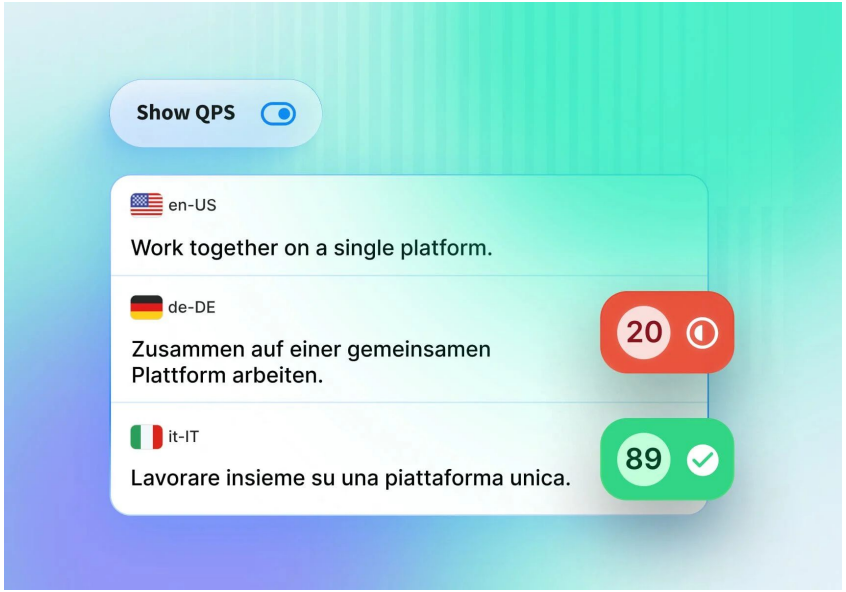

Our foundational strategy is to **lift the lid on quality outcomes at each step** to better enable customers to make informed decisions about how to optimize their solution.

Automated quality checks on source and translation provide **gating mechanisms to maximize quality and minimize the need for human intervention**. First released in TMS in December, with an updated improved model deployed in April.



# QPS-based Human Post-Editing Optimization

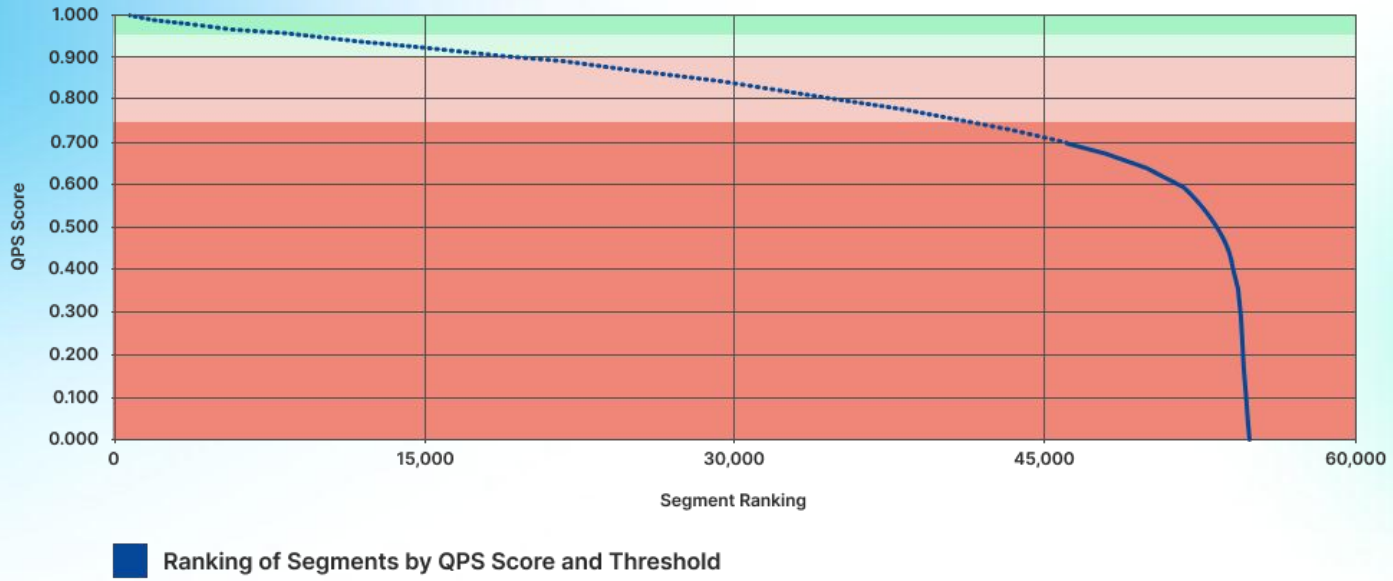
- Implemented inside the Phrase CAT Editor
- Segments that clear the set QPS threshold are blocked from human Post Editing
- Job Profile settings controls QPS Threshold
- Editing is done with full job context and view



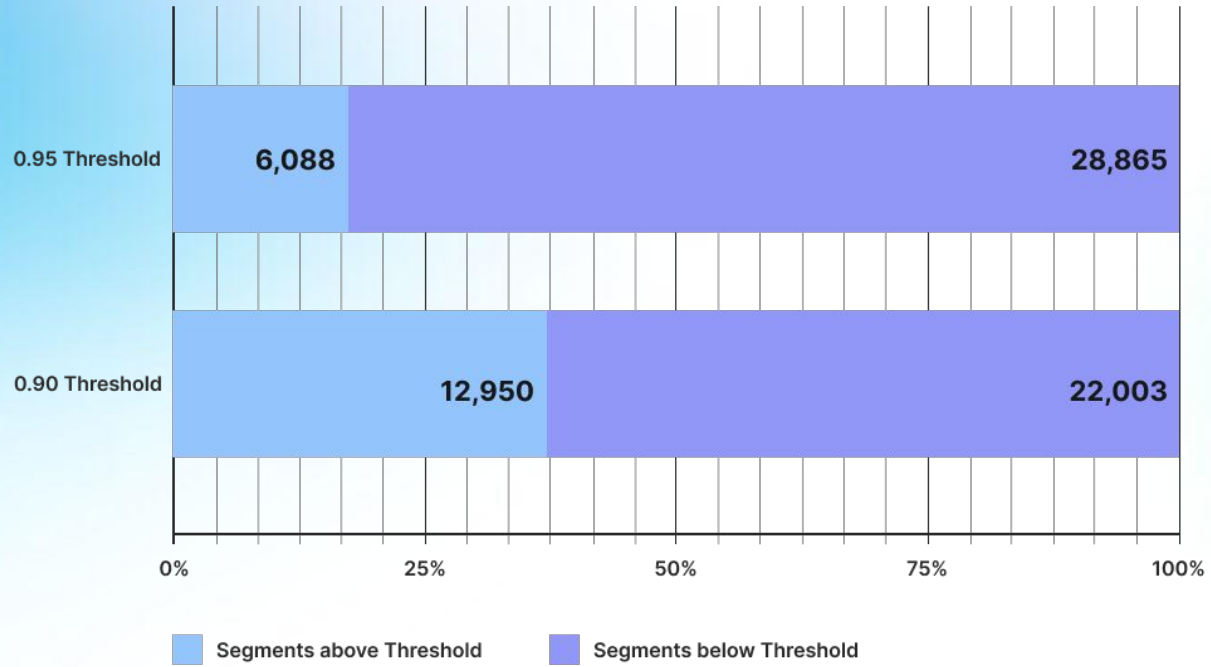
# Modeling QPS-based Quality/Cost Tradeoff

- Utilizes an automated workflow with logic that forgoes human editing based on QPS thresholds
  - Jobs that clear the set QPS threshold are not sent to human PE
  - Segments that clear the set QPS threshold are blocked from human PE
- Consider a base cost of $0.12 per word for human PE
- Calculates cost savings per QPS threshold for all segments/jobs not touched by PE
- Establishes a basis for evaluating QPS that realistically reflects expected Quality/Cost Tradeoff

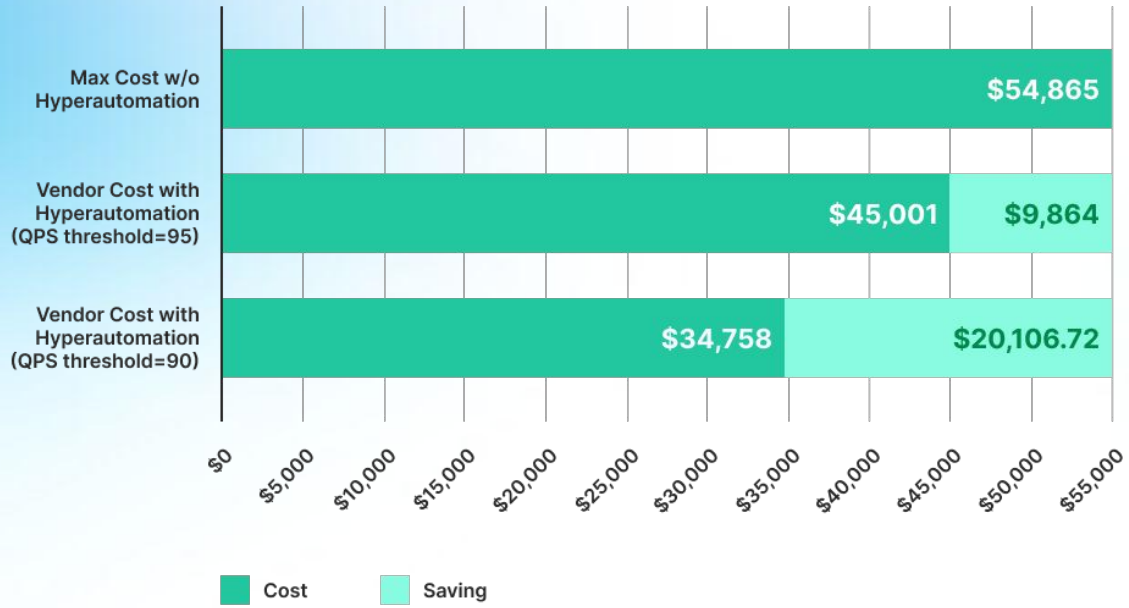
# Segment Quality Ranking by QPS Score



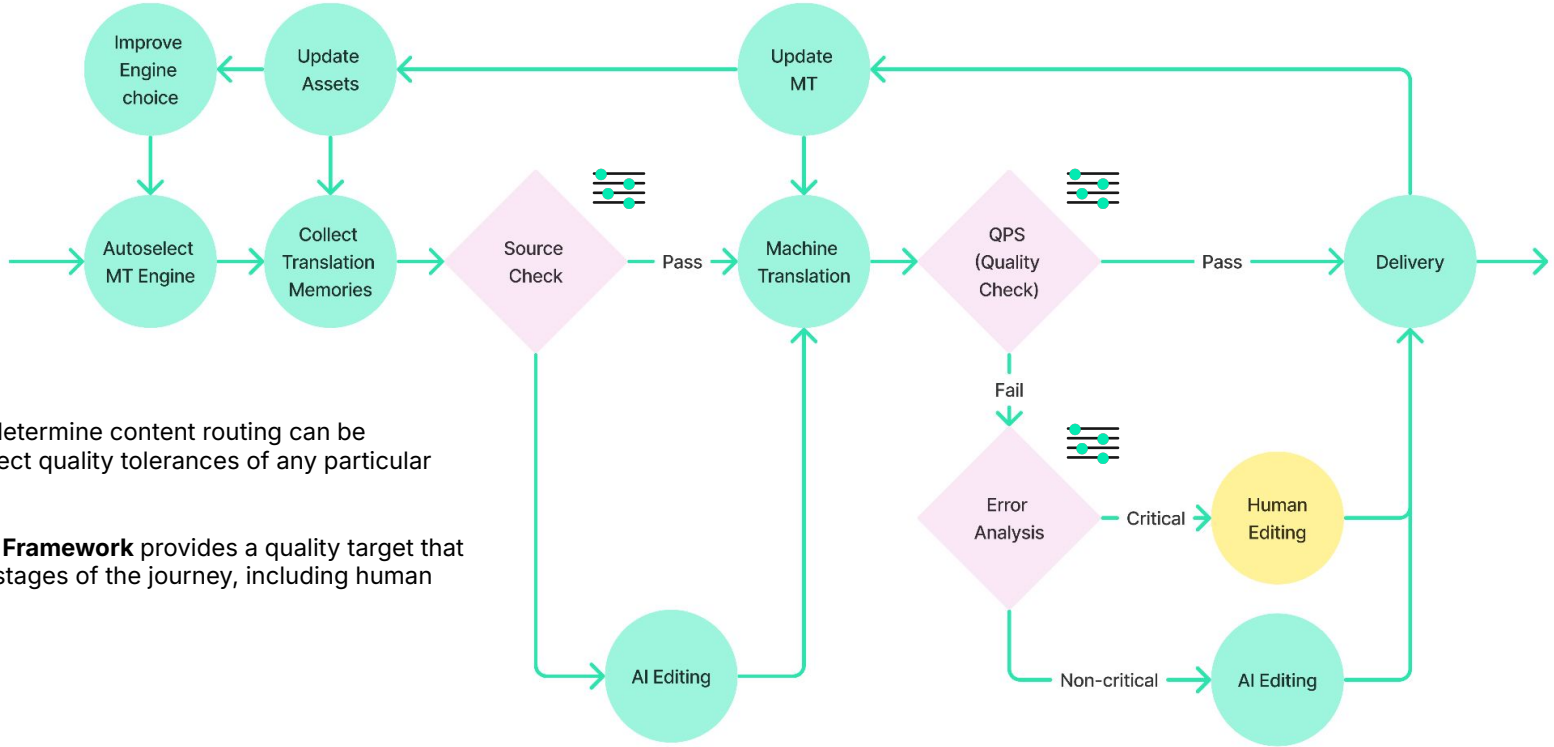
# Segment Quality Distribution by QPS Threshold



# Quality/Cost Tradeoff based on QPS Threshold



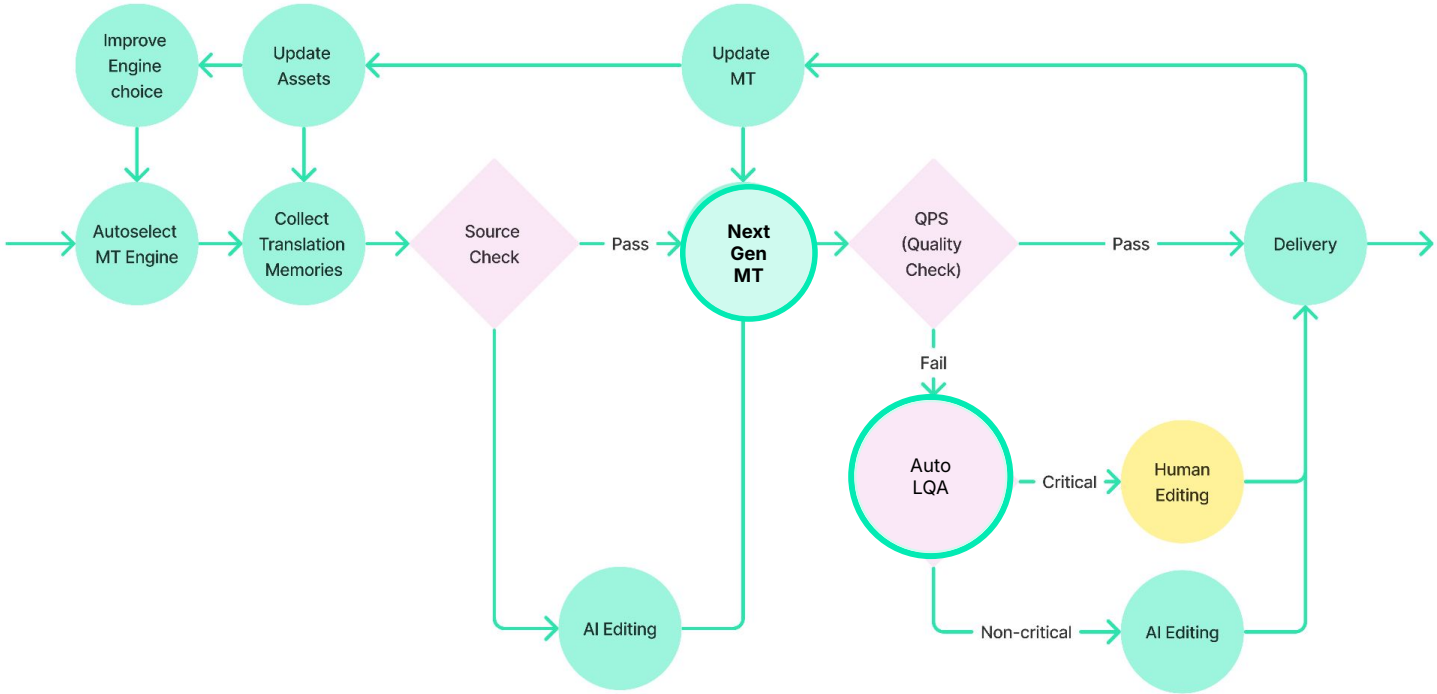
# Hyper-Automated Workflows, powered by TQT



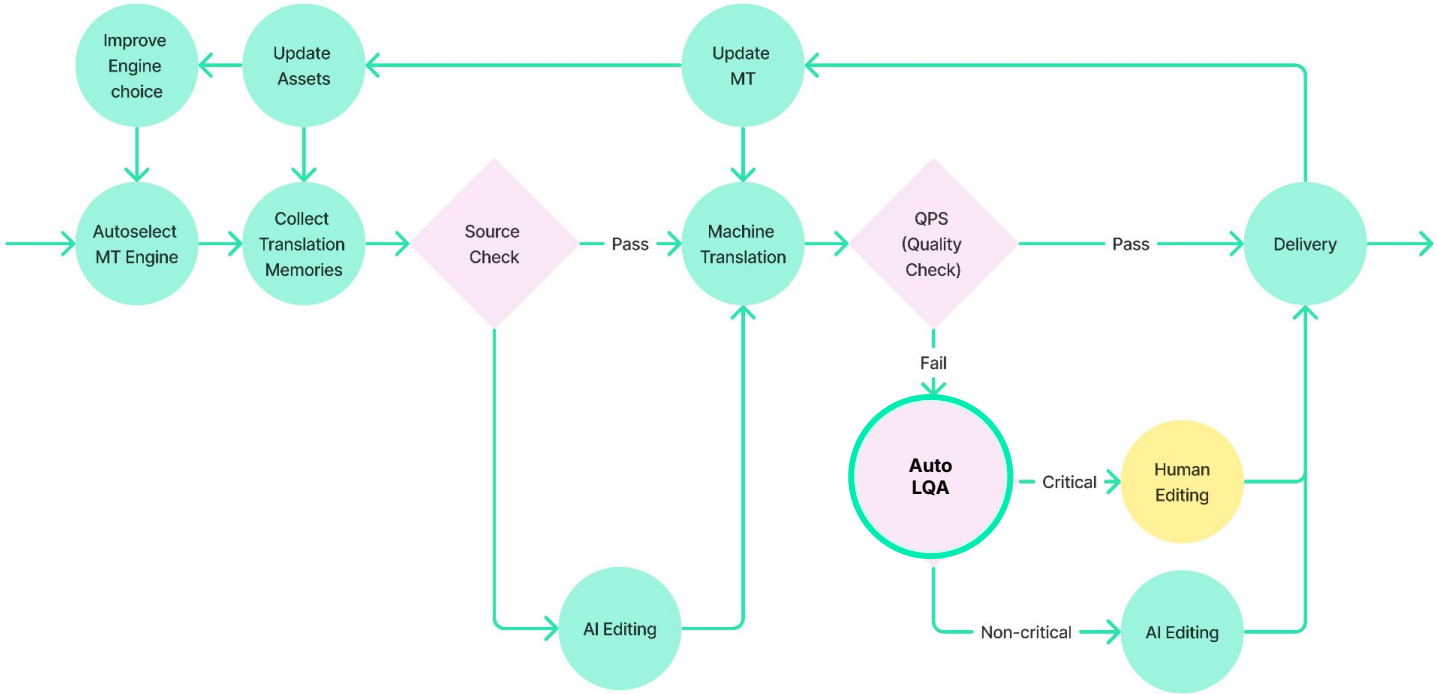

**Thresholds** that determine content routing can be configured to reflect quality tolerances of any particular use case.

A **Unified Quality Framework** provides a quality target that is reflected at all stages of the journey, including human review.

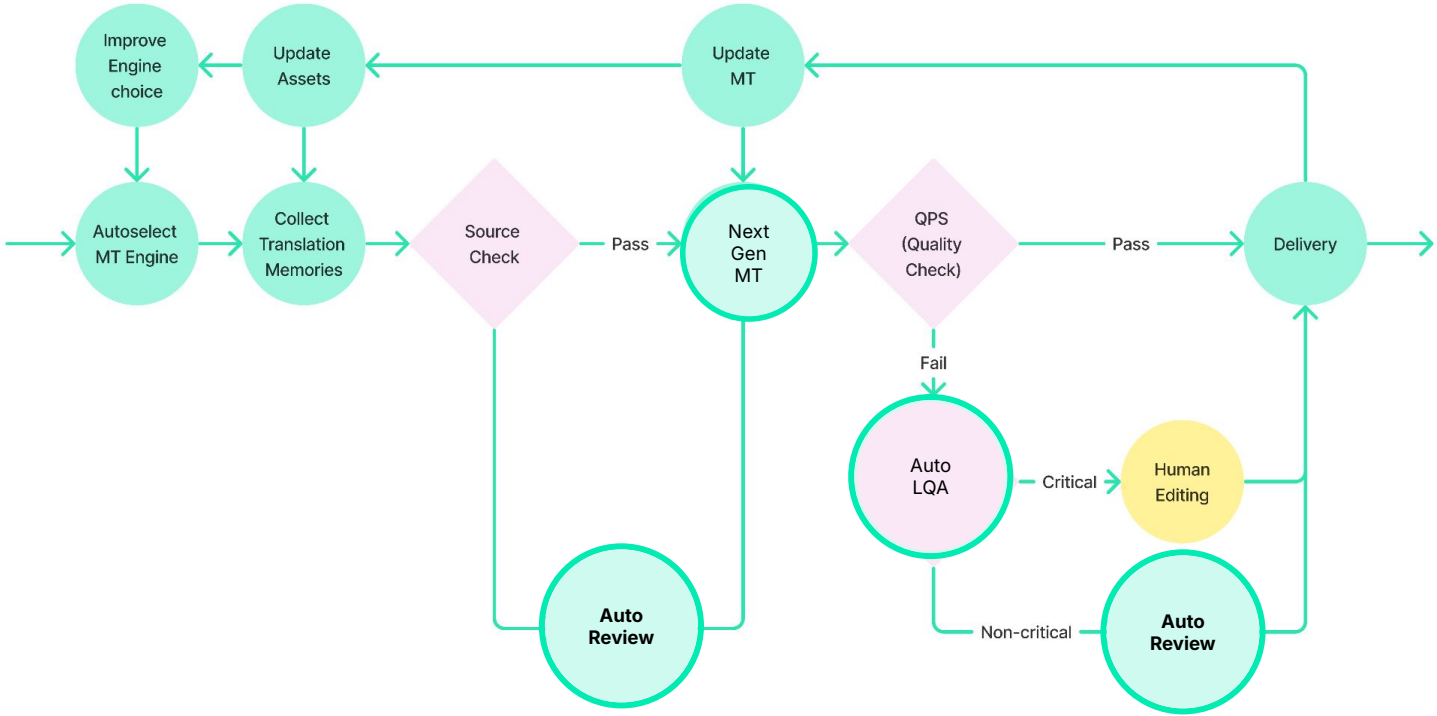
# LLM-based Capabilities Throughout the Phrase Platform



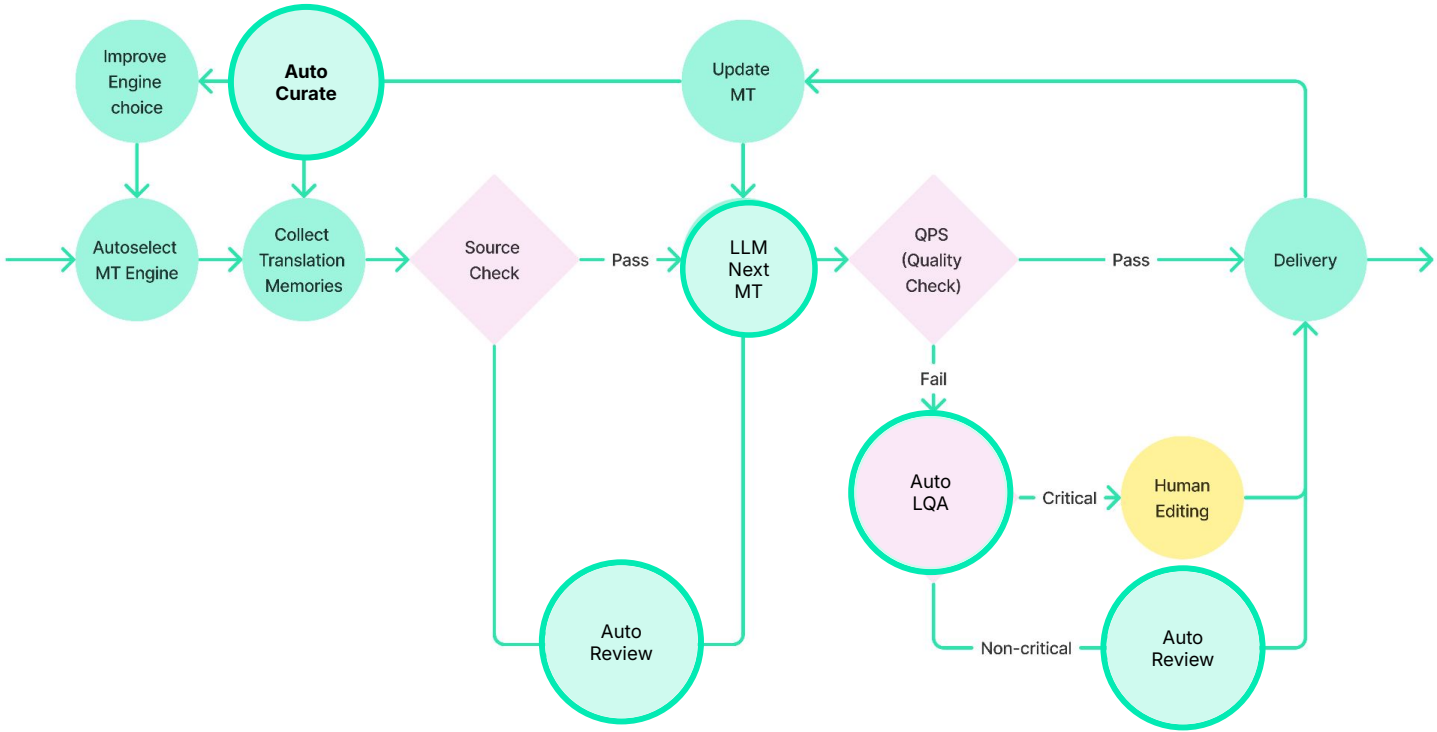
# LLM-based Capabilities Throughout the Phrase Platform



# LLM-based Capabilities Throughout the Phrase Platform



# LLM-based Capabilities Throughout the Phrase Platform



# Phrase Language AI

Game-changing translation technology

## MT AutoSelect

Intelligently selects the most appropriate translation engine for each piece of your content.

Dynamically updates selection mechanism using post-editing insights.

## NextMT

A high-quality, scalable translation engine that draws on language assets like TMs to adapt translation at runtime.

## Next GenMT

Phrase's Generative AI-powered translation engine, with terminology support and advanced tag handling.

Use language assets to customize translation through few-shot prompting (coming December) for superior fluency and accuracy.

# Customer Dynamic Adaptation (RAG) at Phrase

**How/Where we do this today:**

- Our neural **“non-LLM” bilingual and multilingual NextMT models** already implement a weaker form of RAG with Glossary terms and TM fuzzy-matches
- Our **new LLM-based NextGenMT engine** implements RAG with “few-shot” prompt augmentation.
- We currently use the existing TMS Translation Memory fuzzy-match retrieval module to also do example retrieval for MT for both solutions.
- We will explore replacing this with vector-space neural-based retrieval.

**How/Where we will do this down the road:**

- **We expect RAG to become the primary way we perform all forms of customization and personalization in all of our AI models and components,** including:
  - **QPS and Auto LQA:** detecting violations of customer-specific error profiles and style preferences
  - **Auto Review:** adapting target text to specific personalized preferences based on examples.
  - **Auto Curate:** extract customer-specific terminology and curate TMs to specific preferences and requirements.

# NextGenMT: Phrase's LLM-based MT System

- LLM-based MT system based on prompt optimized for performing MT with segment-level input augmented with retrieved "few-shot" examples
- Includes terminology and tag handling capabilities (similar to our neural MT systems)
- No model fine-tuning or any other static customization or training
- Translation performed by OpenAI's GPT models via API calls

**Benchmarking Test Suite:**

- 5 customers with extensive TMs (well suited for few-shot prompting)
- 15 Language Pairs (EN to AR,DA,DE, EL, ES, FR, HU, ID, IT, KO, NB, PL, PT, TH and TR)
- Multiple domains and content-types
- 63 combined data sets
- 25,000 segments in total

# NextGenMT: Phrase's LLM-based MT System

| Engine | Shots | COMET | QPS | BLEU | Chrf3 | TER |
|---|---|---|---|---|---|---|
| **Google MT** | | | | | | |
| **google** | **0** | **89.25** | **81.87** | **50.35** | **68.37** | **35.16** |
| **NextMT** | | | | | | |
| nextMT_bilingual | 0 | 87.78 | 82.67 | 49.42 | 67.53 | 35.98 |
| **nextMT_bilingual** | **1 topN** | **89.45** | **83.27** | **61.79** | **75.26** | **28.06** |
| **NextGenMT with gpt-3.5-turbo-0125** | | | | | | |
| gpt-3.5-turbo-0125 | 0 | 82.43 | 74.64 | 43.09 | 55.52 | 48.47 |
| gpt-3.5-turbo-0125 | 1 rand | 87.76 | 78.88 | 48.05 | 61.95 | 40.75 |
| gpt-3.5-turbo-0125 | 1 topN | 93.07 | 81.7 | 76.52 | 85.32 | 17.75 |
| **gpt-3.5-turbo-0125** | **3 topN** | **94.26** | **82.22** | **79.84** | **87.66** | **14.33** |

# NextGenMT: Phrase's LLM-based MT System

| Engine | Shots | COMET | QPS | BLEU | Chrf3 | TER |
|---|---|---|---|---|---|---|
| **NextGenMT with gpt-3.5-turbo-0125** | | | | | | |
| gpt-3.5-turbo-0125 | 0 | 82.43 | 74.64 | 43.09 | 55.52 | 48.47 |
| gpt-3.5-turbo-0125 | 1 rand | 87.76 | 78.88 | 48.05 | 61.95 | 40.75 |
| gpt-3.5-turbo-0125 | 1 topN | 93.07 | 81.7 | 76.52 | 85.32 | 17.75 |
| **gpt-3.5-turbo-0125** | **3 topN** | **94.26** | **82.22** | **79.84** | **87.66** | **14.33** |
| **NextGenMT with gpt-4-0613** | | | | | | |
| gpt-4-0613 | 0 | 89.02 | 79.84 | 50.97 | 65.22 | 35.93 |
| gpt-4-0613 | 1 rand | 89.67 | 80.41 | 51.74 | 66.18 | 34.51 |
| gpt-4-0613 | 1 topN | 94.17 | 82.49 | 81.06 | 88.18 | 13.51 |
| **gpt-4-0613** | **3 topN** | **94.79** | **82.72** | **83.73** | **90.06** | **11.35** |

# NextGenMT: Phrase's LLM-based MT System

| Engine | Shots | COMET | QPS | BLEU | Chrf3 | TER |
|---|---|---|---|---|---|---|
| **NextGenMT with gpt-4-0613** | | | | | | |
| gpt-4-0613 | 0 | 89.02 | 79.84 | 50.97 | 65.22 | 35.93 |
| gpt-4-0613 | 1 rand | 89.67 | 80.41 | 51.74 | 66.18 | 34.51 |
| gpt-4-0613 | 1 topN | 94.17 | 82.49 | 81.06 | 88.18 | 13.51 |
| **gpt-4-0613** | **3 topN** | **94.79** | **82.72** | **83.73** | **90.06** | **11.35** |
| **NextGenMT with gpt-4o-2024-05-13** | | | | | | |
| gpt-4o-2024-05-13 | 0 | 87.35 | 79.24 | 47.27 | 64.89 | 40.54 |
| gpt-4o-2024-05-13 | 1 rand | 90.23 | 81.16 | 54.48 | 68.5 | 32.38 |
| gpt-4o-2024-05-13 | 1 topN | 94.48 | 82.62 | 83.41 | 89.74 | 11.88 |
| **gpt-4o-2024-05-13** | **3 topN** | **95.03** | **82.81** | **85.51** | **91.18** | **10.04** |

# NextGenMT: Phrase's LLM-based MT System

| Engine | Shots | COMET | QPS | BLEU | Chrf3 | TER |
|---|---|---|---|---|---|---|
| **NextGenMT with gpt-4o-2024-05-13** | | | | | | |
| gpt-4o-2024-05-13 | 0 | 87.35 | 79.24 | 47.27 | 64.89 | 40.54 |
| gpt-4o-2024-05-13 | 1 rand | 90.23 | 81.16 | 54.48 | 68.5 | 32.38 |
| gpt-4o-2024-05-13 | 1 topN | 94.48 | 82.62 | 83.41 | 89.74 | 11.88 |
| **gpt-4o-2024-05-13** | **3 topN** | **95.03** | **82.81** | **85.51** | **91.18** | **10.04** |
| **NextGenMT with gpt-4o-2024-08-06** | | | | | | |
| gpt-4o-2024-08-06 | 0 | 90.01 | 81.27 | 53.39 | 67.86 | 33.13 |
| gpt-4o-2024-08-06 | 1 rand | 90.27 | 81.31 | 54.31 | 68.56 | 32.57 |
| gpt-4o-2024-08-06 | 1 topN | 94.38 | 82.57 | 82.49 | 89.1 | 12.41 |
| **gpt-4o-2024-08-06** | **3 topN** | **95.03** | **82.64** | **84.77** | **90.86** | **10.29** |

# NextGenMT: Phrase's LLM-based MT System

| Engine | Shots | COMET | QPS | BLEU | Chrf3 | TER |
|---|---|---|---|---|---|---|
| **NextGenMT with gpt-4o-2024-08-06** | | | | | | |
| gpt-4o-2024-08-06 | 0 | 90.01 | 81.27 | 53.39 | 67.86 | 33.13 |
| gpt-4o-2024-08-06 | 1 rand | 90.27 | 81.31 | 54.31 | 68.56 | 32.57 |
| gpt-4o-2024-08-06 | 1 topN | 94.38 | 82.57 | 82.49 | 89.1 | 12.41 |
| **gpt-4o-2024-08-06** | **3 topN** | **95.03** | **82.64** | **84.77** | **90.86** | **10.29** |
| **NextGenMT with gpt-4o-mini-2024-07-18** | | | | | | |
| gpt-4o-mini-2024-07-18 | 0 | 89.25 | 80.4 | 51.33 | 66.12 | 35.18 |
| gpt-4o-mini-2024-07-18 | 1 rand | 89.57 | 80.63 | 51.33 | 66.31 | 35.58 |
| gpt-4o-mini-2024-07-18 | 1 topN | 94.21 | 82.43 | 81.43 | 88.74 | 13.39 |
| **gpt-4o-mini-2024-07-18** | **3 topN** | **94.75** | **82.71** | **83.38** | **90.38** | **11.49** |

# Contrasive Costs of Running NextGenMT

- Standard commercial cost of Google MT:  $20 per 1M Characters
- Contrastive cost (per 1MC) of NextGenMT with various OpenAI models

| Google | NextMT (bilingual) | GPT-3.5 (turbo-0125) | GPT-4 (0613) | GPT-4o (2024-05-13) | GPT-4o (2024-08-06) | GPT-4o-mini (2024-07-18) |
|---|---|---|---|---|---|---|
| **$20.00** | $1.00 | $2.00 | $40.00 | $20.00 | $11.00 | **$0.66** |

# Auto LQA: LLM-based Automated LQA

**Perform a fully-automated MQM translation quality annotation:**

- Detect and explicitly identify translation errors at the word span level
- Assign an error category to each error according to the MQM typology
- Assign each error a severity
- (Optionally) provide an explanation of the error and a suggested correction
- Standard MQM score is then calculated algorithmically based on the annotation

# Auto LQA: LLM-based Automated LQA

**Applications and value proposition:**

- Fully-automated alternative to very costly Human LQA
- Human-corrected Auto LQA - reduce costs, improve consistency of Human LQA
- Guide and optimize the work of MT Post-editors within the CAT Editor
- Route low-quality MT for detailed automated error analysis followed by correction
- Ultimately unify with QPS and use Auto LQA MQM scores across the entire platform

# Auto LQA: LLM-based Automated LQA

**How we do it (now):**

- Optimized prompt to an OpenAI LLM, including: the task description, several example annotations and structured output requirements
- Segment-by-segment level input and output
- First productized version (June 2024) uses GPT-4
- Second version (August 2024) reduced false positives and added terminology

**Planned future improvements:**

- Fine-tune the LLM model specifically to perform Auto LQA (i.e. with GPT-4o-mini)
- Include QPS model score within the prompt
- Multi-segment input and output (document-level)
- Customer-specific dynamic adaptation to terminology, error severities and other specifications via larger context prompts and few shot examples

# Auto LQA: LLM-based Automated LQA

**Example:**

| | |
|---|---|
| **Source** | Knit turtleneck sweater made from RWS-certified pure light wool with stocking stitch featuring a ribbed collar, cuffs, and bottom band. |
| **Annotated Target (Auto LQA)** | Jersey de punto de cuello alto en pura lana ligera con certificación RWS realizado en punto liso, presenta cuello, puños y suela de canalé (major, mistranslation). |
| **Annotated Target (Gold)** | Jersey de punto de cuello alto en pura lana ligera con certificación RWS realizado en punto liso, presenta cuello, puños y suela (minor, mistranslation) de canalé. |

The term 'suela de canalé' is incorrect. The correct translation for 'bottom band' in this context should be 'banda inferior'. 'Suela' refers to the bottom part of a shoe, not a sweater.

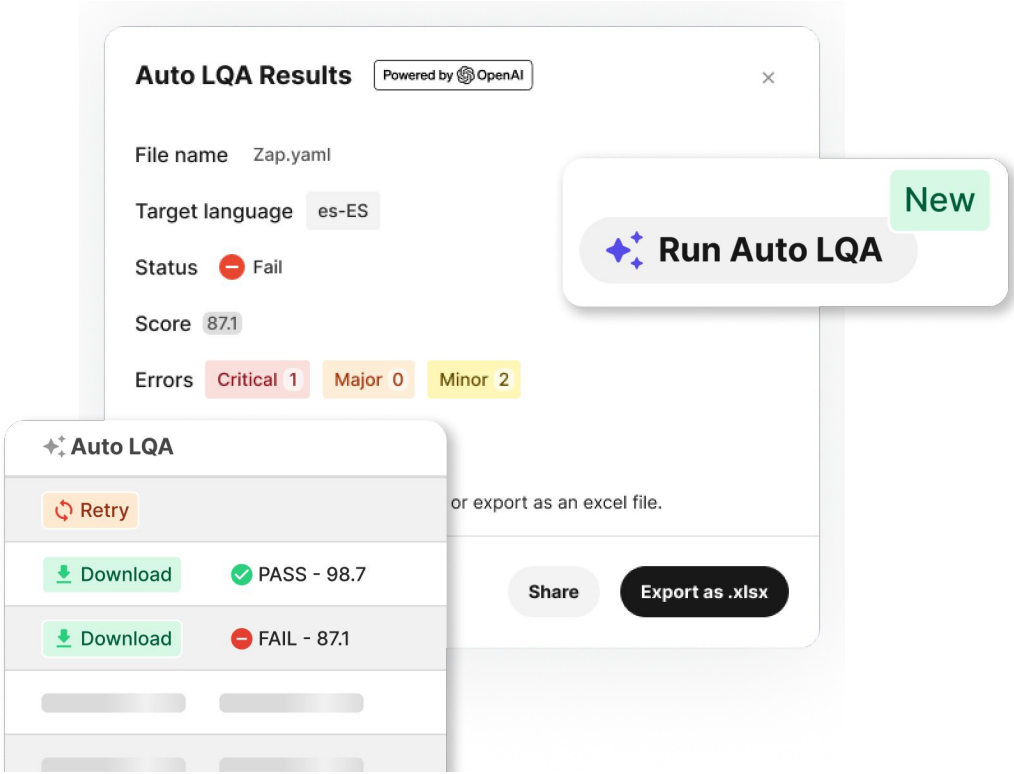
# Auto LQA: LLM-based Automated LQA

- **Evaluation results:**

| model | engine | span_P | minor_R | major_R | #pred_err | FP_ratio | mqm_corr | mae |
|---|---|---|---|---|---|---|---|---|
| **06-24** | gpt-4 | 0.34 | 0.50 | 0.59 | 0.66 | 0.17 | 0.35 | 0.18 |
| **08-24** | gpt-4 | 0.38 | 0.25 | 0.37 | 0.33 | **0.08** | **0.31** | **0.16** |

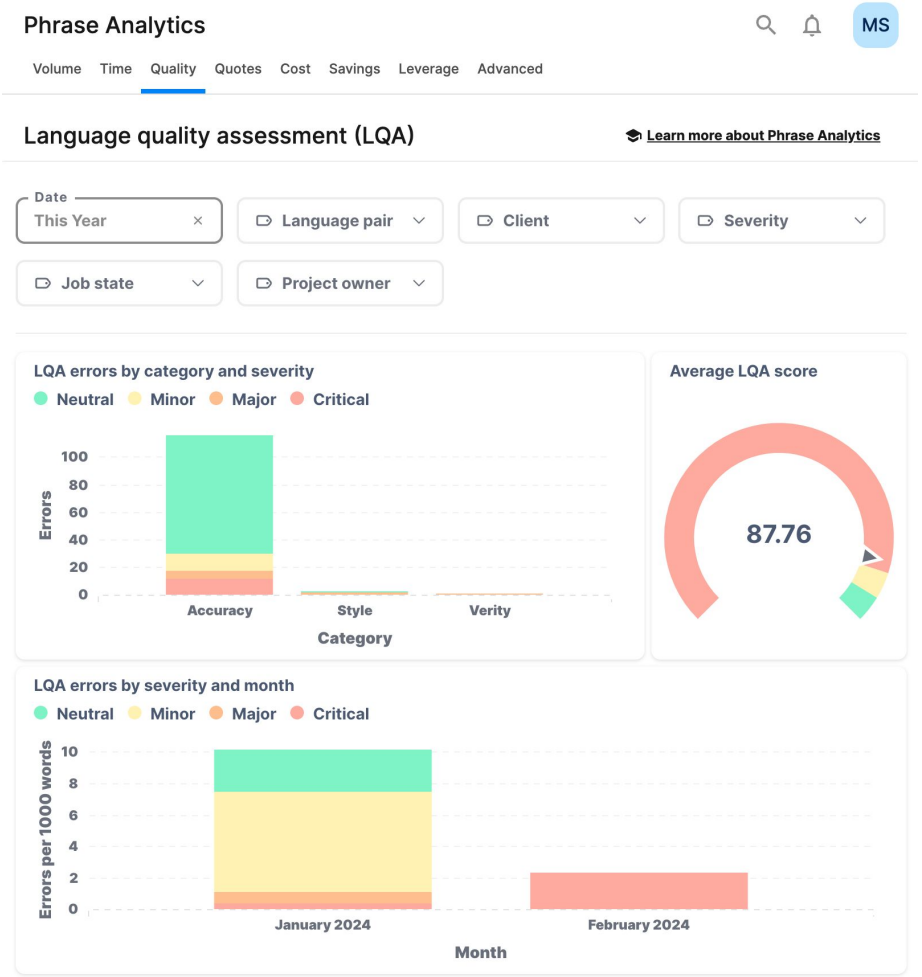
# Auto LQA: Automating Human LQA

- Language Quality Assessment
- LQA is traditionally human-based process and is a major *drain on localization budgets*.
- Auto LQA replaces the human error annotation, powered by OpenAI.
- Re-uses the existing LQA implementation
- Can be launched any time at any workflow step
- Score calculation can be disabled if not needed



# Auto LQA: Analytics

- Auto LQA reuses existing LQA functionality, including the Analytics dashboards
- Users can switch between standard human LQA and the Auto LQA data
- Supports users on the journey to AI tooling for traditional processes.



# Auto Review: LLM-based Job-level Adaptation

**The Problem:**

- Most translation jobs are long-form documents or web pages... but our automated workflows (including MT) operate at the segment-by-segment level
- How can we ensure document-level translation consistency and adherence to customer requirements (terminology, style, formality) without any human review?

**Goal:** Perform a fully-automated document-level review of the translation and correct inconsistencies - similar to Auto PE, but specifically focused on document context

**Applications and value proposition:**

- Fully-automated alternative to costly Human Review
- Apply as an automated step to jobs that clear job-level QPS threshold
- Apply as an automated correction step before or after Auto LQA

# Auto Review: LLM-based Job-level Adaptation

**How we do it (now):**

- Optimized prompt to an OpenAI LLM, including: the task description, customer profile requirements (terminology, formality), and structured output requirements
- Input consists of multiple segments of source and target translation text
- Output consists of (partially) revised target text (for each segment)
- Pilot version developed and tested with one major enterprise customer
- Pilot version uses GPT-4

**Planned future improvements:**

- Fine-tune the LLM model specifically to perform AutoReview (i.e. with GPT-4o-mini)
- Experiment with longer document contexts (GPT-4o-long)
- Customer-specific dynamic adaptation to other specifications via large context prompts and few shot examples
- Interactivity and iteration - allow the customer to augment the prompt and iterate

# Auto Review: LLM-based Job-level Adaptation

## Example:

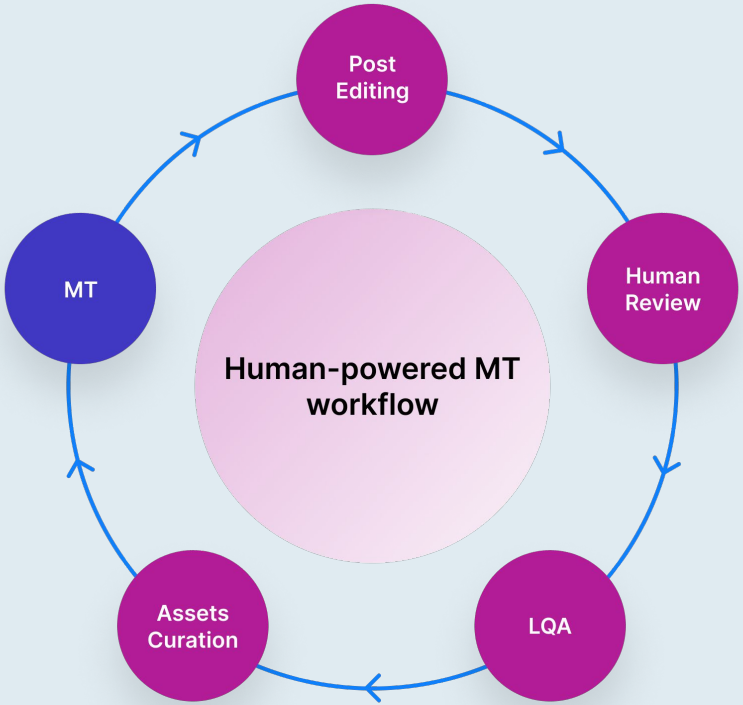
| Source | MT | AutoReview |
| --- | --- | --- |
| [TAG] Well, these questions really show [TAG] how different we all are, don't they? | [TAG] Bueno, estas preguntas realmente demuestran a los [TAG] lo diferentes que somos, ¿no? | [TAG] Bueno, estas preguntas realmente demuestran [TAG] lo diferentes que somos, ¿no? |

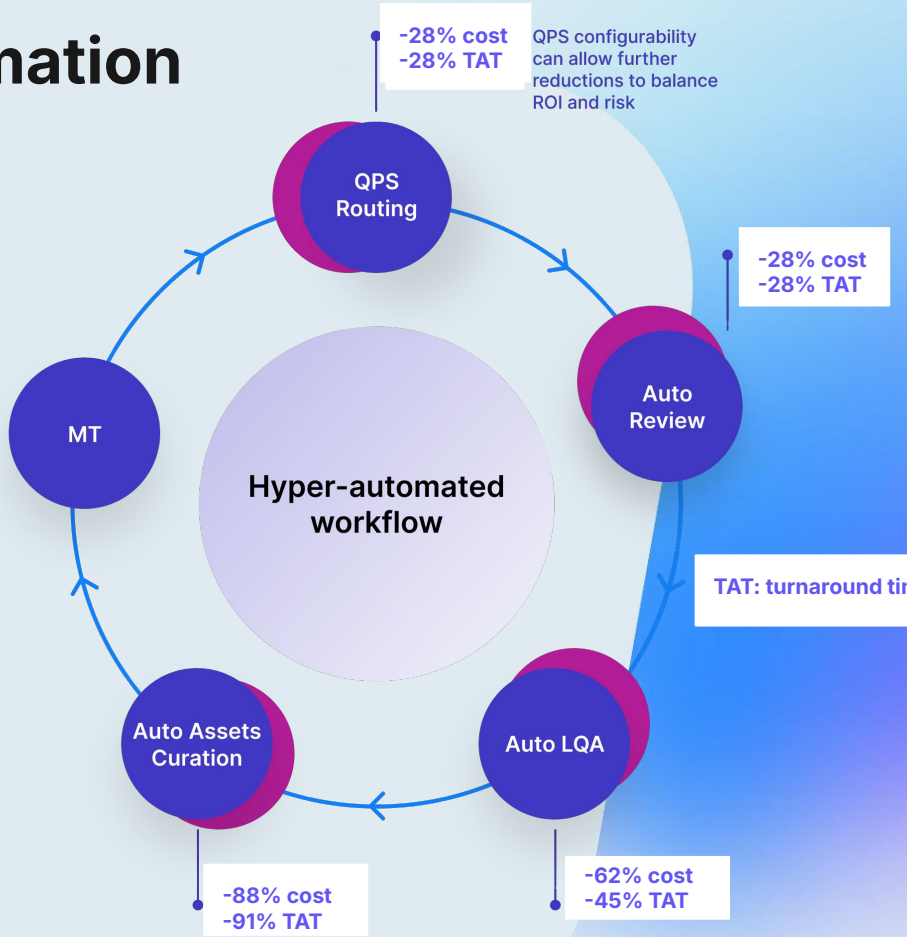
**In a pilot study across multiple languages, linguists noted 87% of content was consistent (19%) or notably improved (68%).**

***"Overall, [AutoReview] improved the MT output drastically…it provided overall a much better text."***

***"[AutoReview] is fixing a lot of hallucinations, omissions and tag placements which is greatly improving the quality."***

# Accelerating Hyper-Automation



# Summary and Takeaway Messages

- Generative AI is driving massive growth in content creation and generation ⇒ massive growth in demand for content localization under fixed budgets
- LLM-based Machine Translation has become very good and continues to improve
- The key to unlocking large-scale enterprise hyper-automation is **complete translation quality visibility**, enabled by **Automated Translation Quality Technology (TQT)**
- Workflows driven by accurate TQT allow enterprises to adopt automated translation at scale while balancing quality risks and cost
- LLM-based advanced technologies such as **Auto LQA** and **Auto Review** can further accelerate automation at scale and close the loop on virtuous feedback cycles driving further efficiency gains

# Summary and Takeaway Messages

**So what's ahead?**

- Consolidation of technologies
  - Unification of QPS and Auto LQA
  - Document-level accurate and consistent MT
  - Consolidation of Auto LQA and Auto Review functionalities
- Increasing focus on translation hyper-customization and personalization
  - Solving the evaluation and assessment problem for "fit for purpose"
- Accurate and personalized Multilingual Content Generation (MCG)

# Any Questions?

## Let’s talk.

Email: alon.lavie@phrase.com


Phrase
Unlock Language. Unlock Opportunity.