From Quality Metrics to Quality Intelligence: How AI Transforms Translation QA into Predictive Insight

José Luis Alcañiz, EMEA & APAC Head of Account Management, Thermo Fisher Scientific Language Services

Objective

Why Rethink Translation Quality Assessment?

Translation Quality Assessment (QA) in Life Sciences traditionally has been reactive, identifying issues after delivery through manual reviews. This approach provides no opportunity for early intervention and limited visibility into systemic risks.

The objectives of this Pilot research are to:

- **Identify and mitigate potential translation errors proactively** by training an AI agent to analyze linguistic QA data before final delivery
- **Benchmark error identification performance** between a non-trained (generic) AI agent and a **customized, Life Sciences–specific AI agent**
- Evaluate whether AI-driven, predictive QA can complement or replace human review by highlighting **high-risk segments and systemic quality drivers** earlier in the process

Goal is to shift linguistic quality management from retrospective compliance toward **proactive, data-driven risk prevention** in regulated translation workflows.

Reimagining Translation Quality Assessment
to Predict and Prevent Issues
Reactive Review
Post-delivery detection
Limited upstream prevention
Reactive, compliance-driven
Predictive Insight
QA Data Analysis
Pattern and risk signal detection
Targeted, early human validation

Methods

Life Sciences - Customized AI Agent for Predictive QA

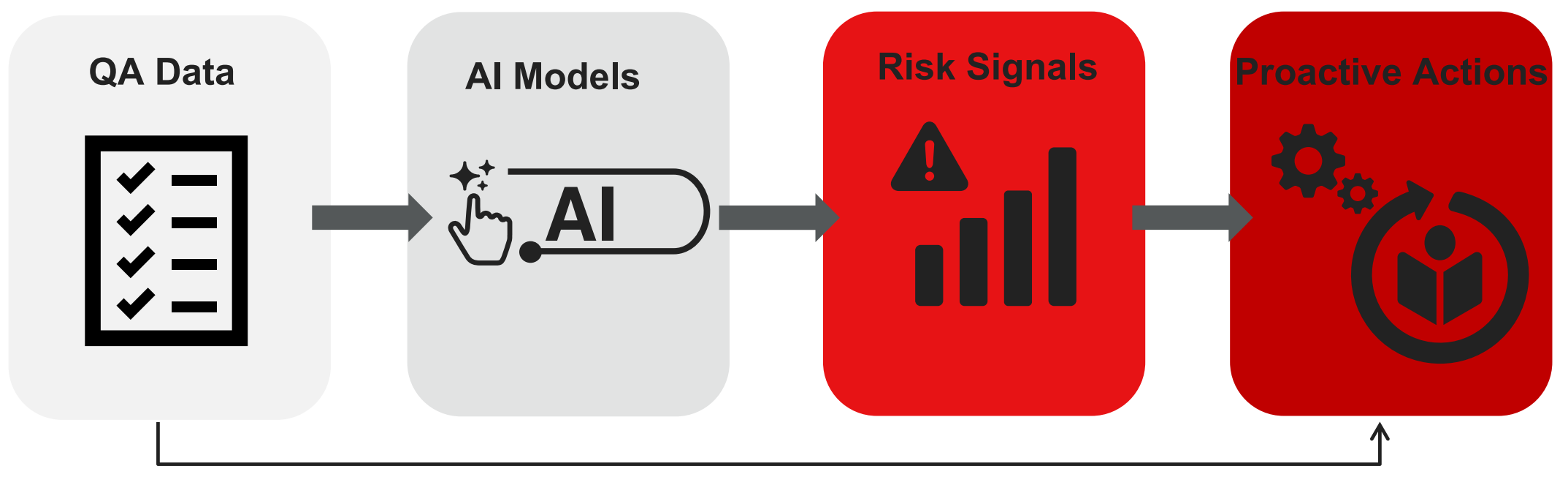
Pilot began with an un-adapted instance of OpenAI's ChatGPT-5.2 to establish a controlled performance baseline. Domain adaptation then performed through structured prompt engineering and integration of Thermo Fisher's proprietary life sciences translation corpora (Korean, Spanish LATAM, German), selected based on the availability of high-volume legacy data to enable targeted model optimization.

Training Inputs

- **Critical clinical trial documents & high-risk sections**
- **Source text analysis** (ambiguous or complex English segments)
- **Translation assets:**
  - Translation Memories (TMs)
  - Glossaries, term bases, and style guides
- **Standardized QA error typologies**, including:
  - Terminology, mistranslation, omission/addition
  - Grammar, syntax, punctuation, readability
  - Formatting, numerical errors, duplication
  - Inconsistencies within and across files
  - Non-adherence to instructions
- **Operational feedback:**
  - Quality Control (QC) rejection data
  - Recurrent reviewer and stakeholder feedback
  - Historical QA findings and trend data

Prompt Engineering & AI Configuration

- **Custom prompts** designed, tested, and refined for regulated workflows
- Performance benchmarked against the **non-trained (generic) AI agent**
- Error identification rates compared to assess **predictive QA capability**



Results

Pilot Results

Initial Pilot demonstrated weak performance for the untrained AI, with modest improvement as AI models were trained and customized.

- **Baseline Model:**

Untrained AI agent identified only **36.0% of translation errors** detected by human review

- **Trained Model:**

Following several rounds of language-specific tuning, the AI agent identified **42.7% of errors** in a second-pass review, meaning **57.3% of known errors were missed**.

- **Key insights:**

Error detection improved to some degree for:

- Terminology inconsistencies
- Repetitive structural errors
- High-frequency, low-visibility issues
- However, we also noted 5.3% of issues in the Baseline Model were subsequently missed in the Trained Model because its detection pattern changed. Had the trained AI "remembered" earlier findings, the final score would have been **48.0%**.

	Introduced Errors	Errors Detected – *Untrained AI*	Errors (Re)Detected – *Trained AI*	Additional Errors Detected – *Trained AI*
Korean	25	5 (20%)	4 (16%)	5 (20%)
Spanish LATAM	25	10 (40%)	9 (36%)	1 (4%)
German	25	12 (48%)	10 (40%)	3 (12%)
Average	75	27 (36%)	23 (30.7%)	9 (12%)

Conclusions

AI agents built on Large Language Models (LLMs) are touted to revolutionize translation, improving speed, quality, and costs. When translation accuracy is critical – for regulated and/or legally binding content, this Pilot study suggests that LLM-based AI for proactive translation quality review does not approach benchmarks set by current ISO 17100 standards, missing over 50% of known errors, even after extensive training and prompt engineering.

- **Limits of Our Approach**
  - Limited Pilot language choices and datasets, namely regulated clinical trial materials
  - We did not capture data on false positives, only detected or missed known errors (true positives)
- **Next Steps**
  - Continue to refine & re-evaluate prompts to drive better, more predictive, results if possible
  - Evaluate other LLM Models (Anthropic, etc.) & additional source/target languages
- **Conclusion**
  - We remain hopeful for LLM-based AI's predictive usefulness for translation QA. However, this Pilot suggests much work is needed in the basic LLM capabilities as well as better training and prompt engineering.

Science at a scan
Scan the QR code on the right with your mobile device to download this and many more scientific posters.

Learn more at ppd.com/our-solutions/clinical/ppd-translate-language-services/