

The State of Translation Automation 2025

An independent multi-domain evaluation of
Machine Translation engines, Large Language Models, and AI Agents

46 MT Engines
and LLMs

11 Language
Pairs

(Previously “State of Machine Translation”)

Disclaimer

April 1—May 20, 2025

The MT systems and LLMs covered in this report were accessed between April 1 and May 20, 2025. Some systems may have been updated since that time period.

Automatic and human evaluation

We employ both automatic and human reference-less evaluations based on the MQM framework, adapted to standard translation requirements: general translation accuracy, terminology compliance, tone of voice, tag handling, and full-text consistency.

When evaluation requirements are different, the assessment must be adjusted accordingly.

In the present study, we use human evaluation to examine the accuracy of automatic LQA. In most practical applications, however, automatic evaluation alone is sufficient.

Data limitations

The evaluation used HTML-tagged and plain text data. Results can vary for other types of markup. We did translation and evaluation on full texts, so in segment-based solutions result may differ.

Valid for a specific dataset

This report shows how the systems performed only on the data listed [on slides 15-17](#), for translation requirements presented [on slide 20](#). We build multiple solutions for our customers, for various language pairs, domains, and translation requirements, and often make a different choice.

Limited customization

In this report, we evaluated baseline models adapted via terminology injection and prompt-tuning. The introduction of fine-tuning for either NMT or LLM systems could yield a different optimal solution, even under identical language pairs and requirements.

There's no “best” MT system or LLM

MT and LLM performance depends on how similar your data is to the data used to pre-train the models, their architecture, and your quality requirements.

Trademarks

All third-party trademarks, registered trademarks, product names, and company names or logos mentioned in the Report are the property of their respective owners, and the use of such third-party trademarks inures to the benefit of each owner. The use of such third-party trademarks is intended to describe the third-party goods or services and does not constitute an affiliation by Intento and its licensors with such company or an endorsement or approval by such company of Intento or its licensors or their respective products or services.

Executive Summary

46 Machine Translation (MT) Engines and Large Language Models (LLMs) evaluated

11 Language pairs

English	→	Spanish*	Ukrainian
		French*	Korean
		Italian	Japanese
		Portuguese*	Chinese*
		German	Arabic
		Dutch	

196K
unique language pairs

+5K
compared to 2024

549
unique languages

Language expansion across all MT engines

* Spanish (LA), French (European), Portuguese (Brazilian), Chinese (Simplified).

14 MT and LLM solutions show the best results between all language pairs

- AI Anthropic Claude Opus 4
- AI Anthropic Claude Sonnet 3.7
- Cohere Command A
- DeepL next-gen
- DeepSeek-V3
- Google Gemini 2.5 Flash
- Google Gemini 2.5 Pro
- Google Gemma 3
- Google NMT
- OpenAI GPT-4.1
- OpenAI o3
- OpenAI o4-mini
- Translated Lara
- Multi-Agent Solution

In this report, we significantly raised the bar for automatic translation. No more sentence-level, reference-based 0–1 similarity scores. We evaluate quality of full-text translation against a specific set of requirements (general translation quality, consistency, terminology, tone of voice, and tag handling).

We **evaluated 80 solutions**, including baseline models, models with light-weight customization (glossaries and tone of voice), human translation, and a multi-agent solution. **LLMs now represent 89% of top performers** (vs 55% in 2024). Multi-agent and customized solutions dominate quality rankings because they adapt better to specific requirements.

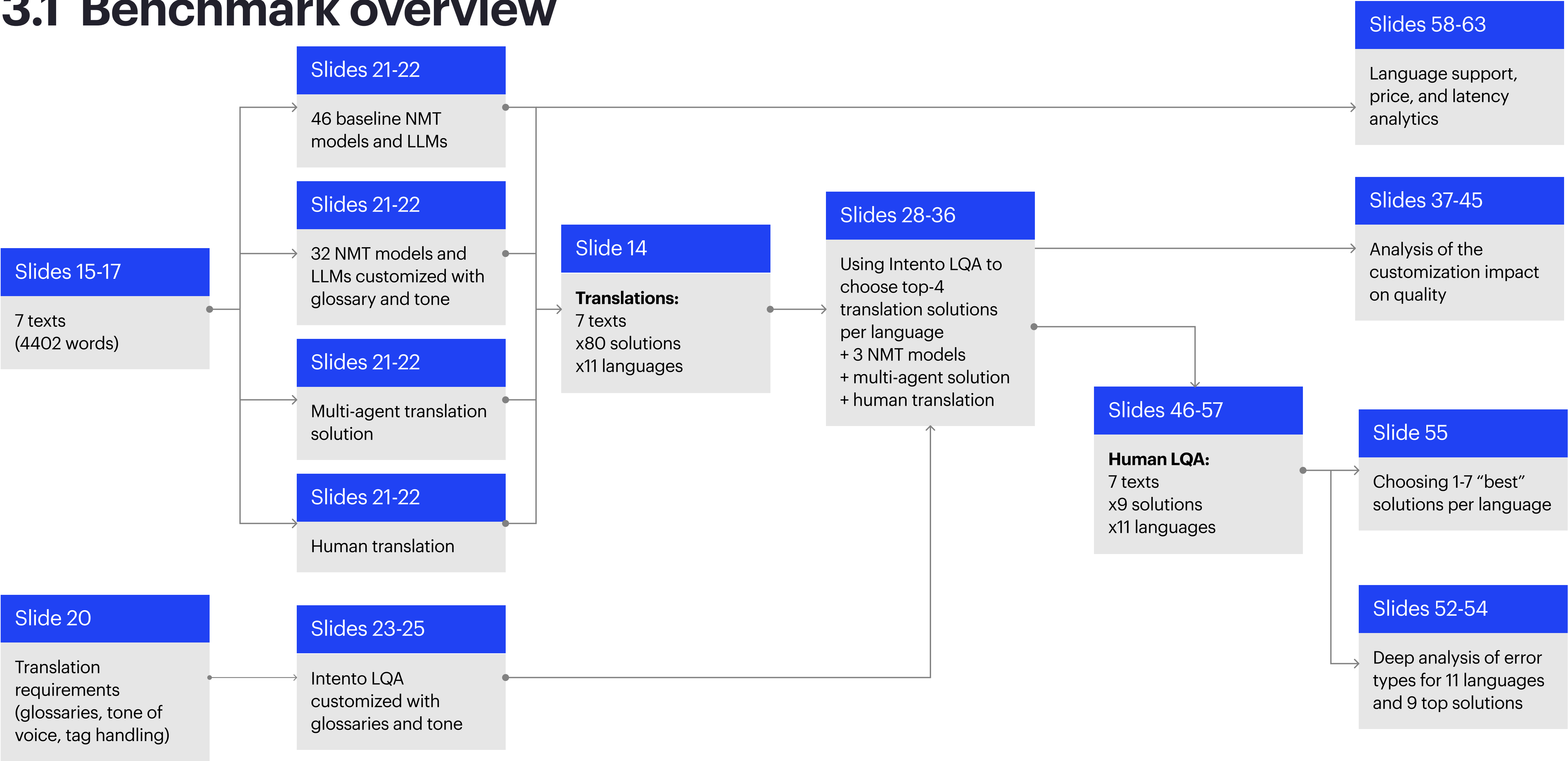
Requirements-based customization is now essential: MT engines and LLMs with glossaries, prompt engineering, and RAG **reduce the amount of errors by 80-90%**. However, some models deteriorate after such customizations. Both automatic and human evaluation cannot distinguish top solutions from human translation. Baseline models fail professional standards with multiple errors per text.

Translation-focused LLMs (**Lara**, **DeepL next-gen**, some **Tower** models) are now achieving latency and price ranges competitive with NMT, eliminating some of the historic cost-quality trade-offs.

From Machine Translation to Translation Automation

- Previously, our NMT and LLM translation quality assessments focused on comparative performance across domains using syntactic and semantic similarity reference-based scores. While these scores indicated relative superiority, they provided no insight into what the lack in quality represented—critical errors or negligible variations.
- This year marks a transition from relative to absolute quality measurement. Rather than determining which system performs better, we now evaluate whether translations meet specific, predefined requirements that reflect real-world translation needs.
- Our current approach evaluates five essential dimensions: general translation quality, terminology accuracy, tone appropriateness, technical precision, and consistency in full text translation. This framework distinguishes between objective quality standards and use-case-specific preferential requirements.
- This methodology transformation enables us to move beyond abstract performance scores to understand precisely how translations succeed or fail against specific requirements—a critical step in advancing from basic machine translation to reliable [translation automation](#).

3.1 Benchmark overview



About Intento

Intento automates translation and localization for global companies using AI agents and machine translation.



We automate localization for global businesses, like Procore and Subway, using AI agents, LLMs and Machine Translation. We deliver consistent and authentic language experience to your customers and coworkers across all touchpoints and software systems in over 500 languages.

How Intento delivers value:

- We build workflows that meet your specific requirements, combining MT customization, AI agents, and Intento tools to achieve **high-quality automated translations**.
- With an extensive connector network that **integrates with over 30 business systems**, your high-quality translations consistently reach every touchpoint.
- Our **team of Language & AI engineers** helps you build, configure, and test automatic translation solutions before launch, then maintains them over time to ensure continuous improvement.

Trusted by the global enterprise



We have been evaluating stock Machine Translation models and building multi-engine translation solutions since 2017. In 2018, we began incorporating fine-tuned NMT models into our workflows. By 2023, we had introduced Source Improvement and Automatic Post-Editing steps to address gaps in NMT output.

As of 2025, our focus has shifted to multi-agent automatic translation solutions designed to meet specific translation requirements.

[Book a demo](#)

Intento Language Hub

Machine translation and multilingual AI agents platform for global enterprise companies.

Intento Language Hub, **ISO-27001 certified**, provides enterprise top-tier security for GenAI solutions in high-demand industries.

Our **ISO-9001 certified** experts help set up, maintain, and continuously refine MT and AI models with new data and user feedback.

[Book a demo](#)



Language Hub for

Localization

AI Agents handle complex translation requirements by fine-tuning MT with historical translations, glossary support, and tone of voice control—even with limited training data. We use automatic quality estimation and automatic post-editing to ensure these enhancements meet your specific quality standards. Works with over 15 TMS.



Customer Experience

Empower customer support teams with real-time machine translation for chats and tickets to help customers 24/7 in their native language. Globalize self-service by on-the-fly translation of knowledge bases and community forums, tailored to your data.



Employee Experience

Our Language Hub translates documents, support tickets, and enterprise apps so everyone's on the same page and included. Works as a translation portal and through integrations with ServiceNow, Atlassian, Microsoft, Google, and other tools.



Overview

1. Neural Machine Translation (NMT) Engines and Large Language Models (LLMs)
2. Datasets
3. Evaluation Methodology
4. Evaluation Results
5. Miscellaneous
6. Takeaways

46

Solutions based on Machine Translation and Large Language Models

11

Language Pairs

1. NMT Engines and Large Language Models

1.1 Machine Translation Landscape

1.2 Generative AI Landscape

1.3 Evaluated MT Engines and LLMs

1.1 Machine Translation Landscape

Generic stock models

AISA	DeepL	Globalese	Kakao	LingvaNex	NiuTrans	Phrase	Rozetta	SYSTRAN by CHAPSVISION	Tilmoch	YarakuZen
Alibaba	Devnagri	Google	Kawamura powered by NICT	Microsoft	NTT COTOHA	Process9	RWS	Tartu NLP	TREBE	Youdao
Amazon	eBay	GTCOM	Kingsoft	Mirai	Omniscien	Prompsit	Sarvam	Tarjama	Ubiquis	
AppTek	Elia	IBM	Lesan	ModernMT	Oracle	PROMT	SmartMATE	Tencent	Unbabel	
Baidu	Fujitsu	iFlyTek	Lindat	Naver	PangeaMT	Reverso	Sogou	Tilde	Yandex	

Vertical stock models

Alibaba	PROMT	XL8
Baidu	RoyalFlush	
CloudTranslation	SAP	
Lingua Custodia	SYSTRAN by CHAPSVISION	
Microsoft	Tartu NLP	
NiuTrans	Tarjama	
Omniscien	Ubiquis	

Custom terminology support

Amazon	RWS
Baidu	SYSTRAN
DeepL	Ubiquis
Google	Yandex
IBM	
Microsoft	
Rozetta	

Static domain adaptation

Alibaba	KantanAI
AppTek	Microsoft
Baidu	Omniscien
CloudTranslation	PangeaMT
Globalese	Prompsit
Google	SYSTRAN by CHAPSVISION
IBM	Tilde

Dynamic domain adaptation

Amazon
ModernMT

Standalone commercial products with an API. All product names, trademarks and registered trademarks are property of their respective owners. All company, product and service names used in this material are for identification purposes only. Use of these names, trademarks and brands does not imply endorsement.

1.1 Machine Translation Landscape

Generic Stock Models

Pre-trained models based on data from multiple sources. These models are not pre-adjusted to one particular industry or specialization, such as Legal or Medical translations.

Vertical Stock Models

Pre-trained models, pre-adjusted to one particular industry or specialization, such as Legal or Medical translations.

Custom Terminology Support

Allows users to customize the MT models by applying their own glossaries. Depending on the implementation, terminology can be used while training custom models or for adjusting machine translation results.

Static Domain Adaptation

The baseline MT model can be adjusted using batch training. The training requires a significant amount of data (thousands of parallel segments) and takes time (from hours to days). Once the model is trained, a snapshot of a model is created and does not change after the next batch retraining.

Dynamic Domain Adaptation

The model can be incrementally updated on the fly. The adaptation can be done with as few as a single datapoint and happens in real-time. Typically, there's no snapshot of the baseline model created, making the model performance affected when the baseline model is updated by an MT provider.

Large Language Models

Large Language Models (LLMs) are trained on massive amounts of data to generate text, follow instructions, and answer questions. These models can be used for various tasks such as content creation, sentiment analysis, text summarization, or translation.

1.2 Generative AI Landscape



Cloud Commercial

01.AI ⚡	Anthropic ⚙️ ⚡ 🧠	Cohere ⚙️ ⚡	HuggingFace ⚙️ ⚡	Mistral ⚙️ ⚡ 🧠	Salesforce ⚙️ ⚡ 🧠	Unbabel ⚙️
AI21 ⚙️ ⚡	Baichuan ⚙️ ⚡	Databricks ⚙️ ⚡	IBM ⚙️ ⚡ 🧠	NVIDIA ⚙️ ⚡ 🧠	Snowflake ⚙️	X.ai ⚡ 🧠
Alibaba ⚙️ ⚡ 🧠	Baidu ⚙️ ⚡ 🧠	DeepSeek ⚙️ ⚡ 🧠	Meta AI ⚙️ ⚡ 🧠	OpenAI ⚙️ ⚡ 🧠	Stability AI ⚙️ ⚡	
Amazon AWS ⚙️ ⚡	Cerebras ⚙️ ⚡	Google ⚙️ ⚡ 🧠	Microsoft Azure ⚙️ ⚡ 🧠	Preferred Networks ⚡	Tencent ⚙️ ⚡ 🧠	

Open Commercial

01.AI ⚙️ ⚡	BAAI ⚙️ ⚡	Cohere ⚙️	EU Commission ⚙️ ⚡	LLMZoo ⚙️ ⚡	Preferred Networks ⚙️ ⚡	Tencent ⚙️ 🧠
Ai21 ⚙️ ⚡	Baichuan ⚙️ ⚡	Databricks ⚙️ ⚡	Google ⚙️ ⚡ 🧠	Meta AI ⚙️ ⚡	Salesforce ⚙️ ⚡	TII (UAE) ⚙️ ⚡
Alibaba ⚙️ ⚡ 🧠	BSC ⚙️ ⚡	Deci ⚙️ ⚡	HuggingFace ⚙️ ⚡	Microsoft Azure ⚙️ ⚡ 🧠	Snowflake ⚙️	X.ai ⚙️
AllenAI ⚙️ ⚡	Big Science ⚙️ ⚡	DeepSeek ⚙️ 🧠	IBM ⚙️ ⚡ 🧠	Mistral ⚙️ ⚡ 🧠	Stability AI ⚙️ ⚡	
AMD ⚙️ ⚡	Cerebras ⚙️ ⚡	Eleuther AI ⚙️ ⚡	LLM360 ⚙️ ⚡ 🧠	NVIDIA ⚙️ ⚡ 🧠	StatNLP ⚙️ ⚡	

Open Non-commercial

01.AI ⚙️ ⚡	Alibaba ⚙️ ⚡	AMD ⚙️ ⚡	Cohere ⚙️ ⚡	Lianjiia Tech ⚡ ⚙️	Preferred Networks ⚡	Stability AI ⚙️ ⚡
AI21 ⚙️ ⚡	AllenAI ⚙️ ⚡	BAIR ⚙️	Databricks ⚙️	Microsoft Azure ⚙️ 🧠	Salesforce ⚙️ ⚡ 🧠	Unbabel ⚙️ ⚡
Aleph Alpha ⚙️ ⚡						

- ⚙️ – Customizable
- ⚡ – Flash
- 🧠 – Reasoning

Provider icons show all model types offered by that provider in each category.

All product names, trademarks, and registered trademarks are the property of their respective owners. All company, product, and service names used in this material are for identification purposes only. Use of these names, trademarks, and brands does not imply endorsement.

Updated on August, 2025.
For any revisions, please reach out to us at hello@inten.to.

Remember to verify the license information, as licenses may change.

1.2 Generative AI Landscape

License types

Cloud Commercial

Premium models that are accessible only as managed cloud APIs. A user pays by subscription or metered usage, and they must have an internet (or private-link) connection to the provider. The code and weights are never downloaded.

Open Commercial

Models whose weights can be downloaded and run on the user's hardware or in any cloud for different purposes, research or business, and whose licenses explicitly allow commercial use.

Open Non-commercial

Models released for research, education, or personal projects only. The weights are downloadable, but any revenue-generating or product use is forbidden unless you obtain separate permission.

Model capabilities

Customizable LLMs

Foundation models that can be adapted to a specific domain through fine-tuning, parameter-efficient methods (LoRA, QLoRA, adapters), in addition to the retrieval-augmented generation (RAG). The result behaves like the model mastered your organization's jargon, style, and proprietary knowledge.

Flash LLMs

A product class (popularized by Google's Gemini Flash models) designed for very low latency and high throughput at the cost of some peak accuracy on complex reasoning or coding benchmarks. Comparable offerings may be called "Lite," "Edge," "Distilled," or "Small" by other vendors.

Reasoning LLMs

Large models optimized and often prompt-engineered for multi-step, chain-of-thought reasoning, logical inference, and contextual understanding.

1.3 Evaluated MT Engines and LLMs

Customization options

- None
- ◐ TM
- ◑ Glossary
- Both

MT Engines

 Alibaba E-Commerce MT ○	 Elia MT API ○	 ModernMT Adaptive ●	 Oracle Machine Translation ○	 Tarjama MT API ○	 TREBE Machine Translation API ○
 Amazon Translate ●	 Google Advanced Translation ●	 Naver Papago NMT Commercial ○	 Pangeanic Machine Translation API ○	 TartuNLP Neurotõlge MT ○	 Yandex Translate API ●
 Baidu Translate API ◐	 HiThink RoyalFlash Finance Translation ○	 NiuTrans Translation Cloud Platform ○	 SYSTRAN PNMT ●	 Tencent Cloud TMT API ○	 Youdao Cloud Translation API ○
 DeepL APi ◐	 Microsoft Language Translator ●				

LLMs

 AI21 Jamba Large 1.6 ●	 Amazon Bedrock Nova Pro ●	 Cohere Command A ●	 DeepSeek R1, V3 ●	 Meta AI Llama 4 Maverick, Llama 4 Scout ●	 Translated Lara ●
 Alibaba Cloud Qwen 2.5 72B Instruct Turbo, QwQ-32B ●	 Anthropic Claude 3.7 Sonnet, 3.5 Haiku, Sonnet 4, Opus 4 ●	 DeepL New Gen model ◐	 Google Gemini 2.5 Flash, 2.5 Pro, 2.0 Flash, Gemma 3, Translation LLM ●	 OpenAI GPT-4.1, o3, o4-mini ●	 Widn AI Tower Anthill, Sugarloaf, Vesuvius ●

Large Language Models can be customized with TMs through fine-tuning, RAG, and terminology via prompt engineering

2. Datasets

2.1 Data Description

2.2 Content Samples by Domain

2.1 Data Description

The source data collection and was done by Intento through Intento's partners.

Using full text data instead of segment translation

- Our objective is to raise the bar by moving from segment-by-segment to a full-text translation approach and addressing client requirements and consistency limitations while keeping fine-tuning and RAG from historical translation out of scope.
- We're proceeding without a reference baseline to ensure our approach remains generalizable to avoid overfitting to specific reference standards.
- [Seven texts](#) have been collected with the help of a multinational fast food restaurant franchise and a general management magazine.
- [Six out of seven](#) texts are HTML-tagged while [one](#) does not have any markup as it a transcript of a talk.

The texts represent 5 content domains

- [Legal](#) (Data Privacy & Rights)
- [Human Resources](#) (Workplace Communication, Productivity & Time Management)
- [Health & Safety](#)
- [Educational](#) (Critical Thinking & Problem-Solving)
- [Marketing](#) (in Language Technology & Translation)

2.2 Content Samples by Domain

Marketing

“...Let me be clear about something really important: this technology is designed to elevate the translation profession, not replace it. Think of AI agents as taking over the tedious parts of post-editing, freeing up translators to focus on the more rewarding aspects of their craft...”

Human Resources

“...To help managers and companies identify employees at risk of quitting, we investigated this very question and uncovered a set of behavioral changes exhibited by employees...”

Health & Safety

“...We’ve even done side-by-side tests in highly trafficked areas of items like sofas with and without forever chemicals. Our cleaning staff indicated that those without forever chemicals were easier to clean because stains were no longer getting trapped below the forever chemical coating...”

Educational

“...Situation #2: You’re asked to stay late when you’re about to leave the office for a personal obligation...”

Legal

“...You can ask Us to delete or remove personal data where there is no good reason for Us to continue to process it...”

3. Evaluation Methodology

- 3.1 From similarity-based scoring to actionable translation quality evaluation with Intento LQA
- 3.2 Specific Translation Requirements
- 3.3 Customizing baseline models and comparing them to human and multi-agent translation
- 3.4 Translation Automation with a Multi-agent Solution
- 3.5 Issue Classification and Severity for Intento LQA
- 3.6 Examples of Issue Classification Using Intento LQA
- 3.7 How we Determine Best Solutions
- 3.8 How Do We Choose Solutions for Human Evaluation?

3.1 From similarity-based scoring to actionable translation quality evaluation with Intento LQA

Moving away from legacy scoring built for NMT

- Traditional NMT models can produce a translation draft, but then can't assess or improve the draft. That was historically done by humans working with machines.
- This is reflected in quality evaluation scores built for machine translation quality estimation: they are calculated on text segments and either directly assess how similar the NMT output is to a human translation (reference-based scores such as COMET) or trained on a dataset of such reference translations (reference-less scores such as MTQE).
- However, such scores don't provide insights into what's wrong with translations, which is necessary for both improving translation quality with LLMs and for selecting the best solution under specific enterprise translation requirements.
- Basically, you know that the solution with 0.8 is better than the solution with 0.75, but you don't know what's missing, what prevents it from being a 1.00.

Requirements-based scoring with Intento LQA

- The core advantage of LLMs over NMT is that LLMs can work with full texts and improve specific aspects of translation drafts.
- Using this advantage requires scoring that evaluates translation quality towards specific important translation requirements instead instead of similarity to a reference translation.
- [Intento LQA](#) evaluates the translation quality based on the source text and translation, and it's customized towards specific translation requirements. It uses the [MQM](#) error typology, providing actionable insights into what's wrong with the translation, which can be leveraged by LLM-based Automatic Post-Editing in multi-agent translation solutions.

3.2 Specific Translation Requirements

→ We are working with the following requirements for full text translation in this year's report:

Basic requirements

- General translation quality
- Translation consistency throughout the whole text

Standard requirements

- Formal tone of voice
- Terminology consistency according to the glossaries extracted from human translation
- Correct formatting for HTML content

→ We did not include advanced requirements—such as inclusivity, style, or idiomaticity—in the evaluation. If human evaluators inadvertently reflected such preferences in their reviews, those parts were excluded.

Advanced
requirements

Inclusivity, style,
idiomaticity, etc.
(not included in evaluation)

Standard
requirements

Terminology

Tone of voice:
formal

HTML
formatting

Basic
requirements

General Translation
Quality

Translation
Consistency

3.3 Customizing baseline models and comparing them to human and multi-agent translation

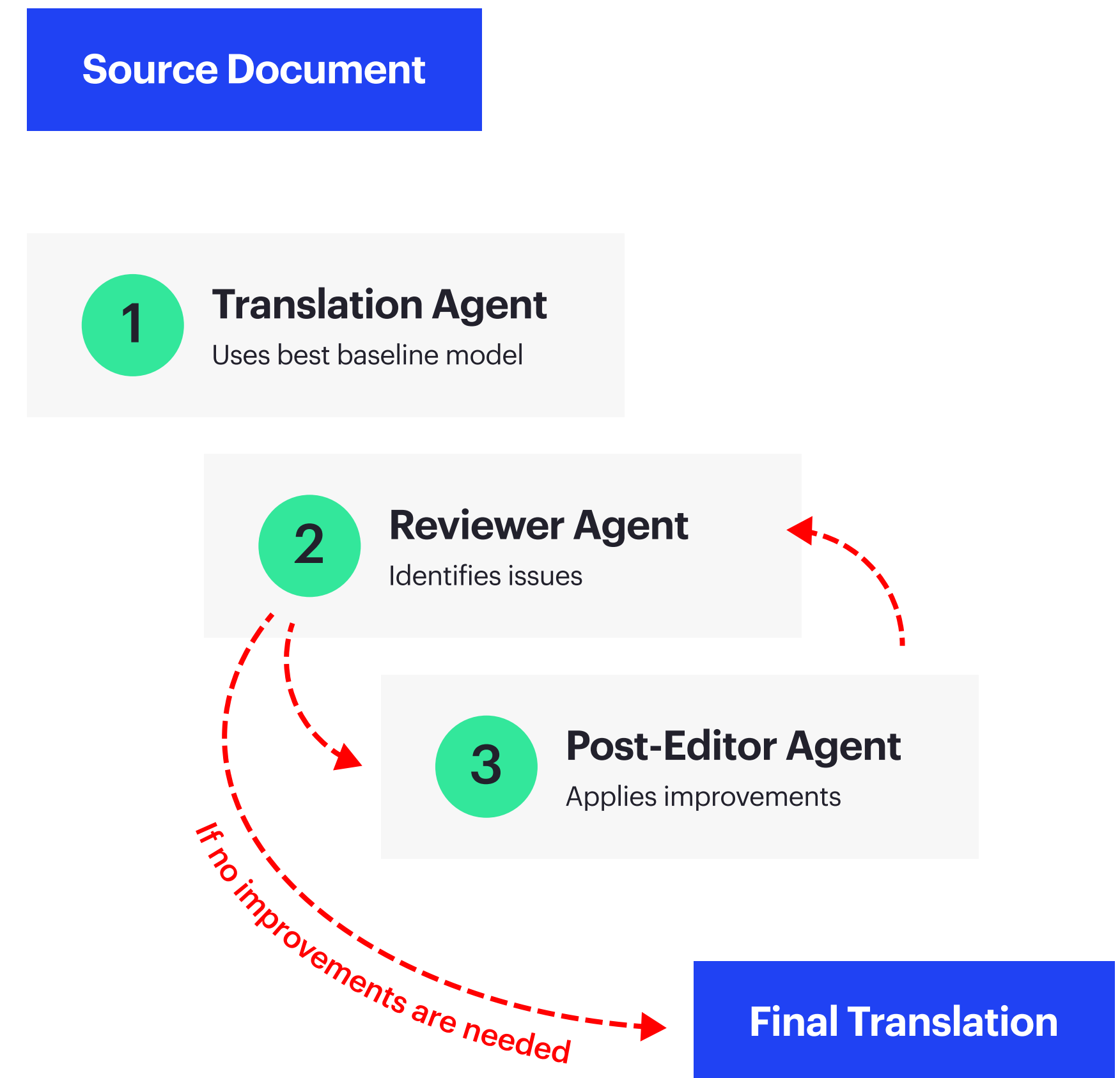
- This year, we evaluate a range of machine translation and large language model solutions, alongside human translation and a multi-agent workflow.
- We compare these approaches to identify which delivers the highest-quality full-text translations while remaining practical to implement at scale.
- Using the human translations, we prepare two **glossaries**: one for common terminology and another for proper nouns.
- Six of the seven texts contain extensive **HTML tagging**, making it essential for the models to preserve consistent formatting throughout.
- We also specify that the **tone of voice** for each text must be **formal**.

The following solutions are evaluated:

-  **Baseline**: Stock NMT models without terminology or tone of voice adjustment and LLMs with basic translation prompts
-  **Customized**: terminology and formal tone of voice applied to NMT where possible or injected into LLM prompts via RAG
-  **Human translation**: human translation via a reputable mid-size LSP with an ordinary quality level (\$0.16-\$0.24/word)
-  **Multi-agent**: Workflow combining the best baseline translation with specialized agents - one for translation review and another one for glossary integration and tone adjustment

3.4 Translation Automation with a Multi-agent Solution

- Three specialized AI agents work with the top baseline translation in each language pair in sequence to ensure translation quality and compliance with client requirements
- **Translation Agent:** translates the full text using the best scoring baseline solution
 - **Reviewer Agent:** identifies issues with tone, terminology, and formatting against predefined standards in the raw machine translation
 - **Post-Editor Agent:** applies targeted improvements based on reviewer feedback without over-editing
 - The same **Reviewer Agent** validates the final quality and confirms that all requirements are met. If post-editing produces no improvements, the original translation is retained



3.5 Issue Classification and Severity for Intento LQA

We have adjusted the latest [MQM](#) error typology for Intento LQA, leaving the issue types that are reliably identified by LLMs (with a limited number of false positives and issue duplicates):

- Accuracy: Addition, Omission, Mistranslation, Over-translation, Under-translation, Untranslated text
- Linguistic conventions: Punctuation, Spelling, Grammar, Inconsistency, Character encoding
- Terminology: Inconsistent use of terminology
- Style: Awkward, Inconsistent, Unidiomatic style
- Design: Local formatting, Markup
- Audience appropriateness: Culture-specific reference

According to the MQM description, we are using the following error severity classification:

- Critical (10-point penalty): Has a significant impact and may cause severe implications
- Major (5-point penalty): Has a considerable impact and may confuse or mislead the reader
- Minor (1-point penalty): Has a slight impact; does not cause loss of meaning nor confuse the reader
- Neutral (0-point penalty): Flag problems that are not considered errors, for example preferred stylistic changes. No penalty associated

[Intento LQA](#) begins at 100 (perfect quality) and deducts penalties for each translation error. The final score is normalized by text length.

3.6 Examples of Issue Classification Using Intento LQA

7-point penalty

Source: “I wish I could have said something better about Take The High Ground because I certainly like its talented cast, its talented director Richard Brooks, even the silly theme by Dimitri Tiomkin and Ned Washington, fresh from their Oscar a year before for High Noon.”

MT: “لأنني بالتأكيد Take The High Ground أتمنى لو كان بإمكانني أن أقول شيئاً أفضل عن أحب طاقمها الموهوب، ومخرجها الموهوب ريتشارد بروكس، وحتى الموضوع السخيف High Noon. لديميتري تيومكين ونيد واشنطن، الجديد من جائزة الأوسكار قبل عام عن

Intento LQA:

- Major error (5-point deduction) - Untranslated text: The title 'Take The High Ground' and 'High Noon' were not translated into Arabic, which may confuse the reader.
- Minor error (1-point deduction) - Grammar: The phrase 'الجديد من جائزة الأوسكار قبل عام' is awkward and ungrammatical. A better translation would be 'الحائز على جائزة الأوسكار قبل عام'.
- Minor error (1-point deduction) - Fluency: The machine translation is generally understandable but lacks the fluency and natural flow of the reference translation.

10-point penalty

Source: “The Commission considers that this will inevitably have led them to temper their competitive behaviour towards each other.”

Machine translation: “欧州委員会は、このことが必然的に相互の競争行動を抑制することにつながったと考えている。”

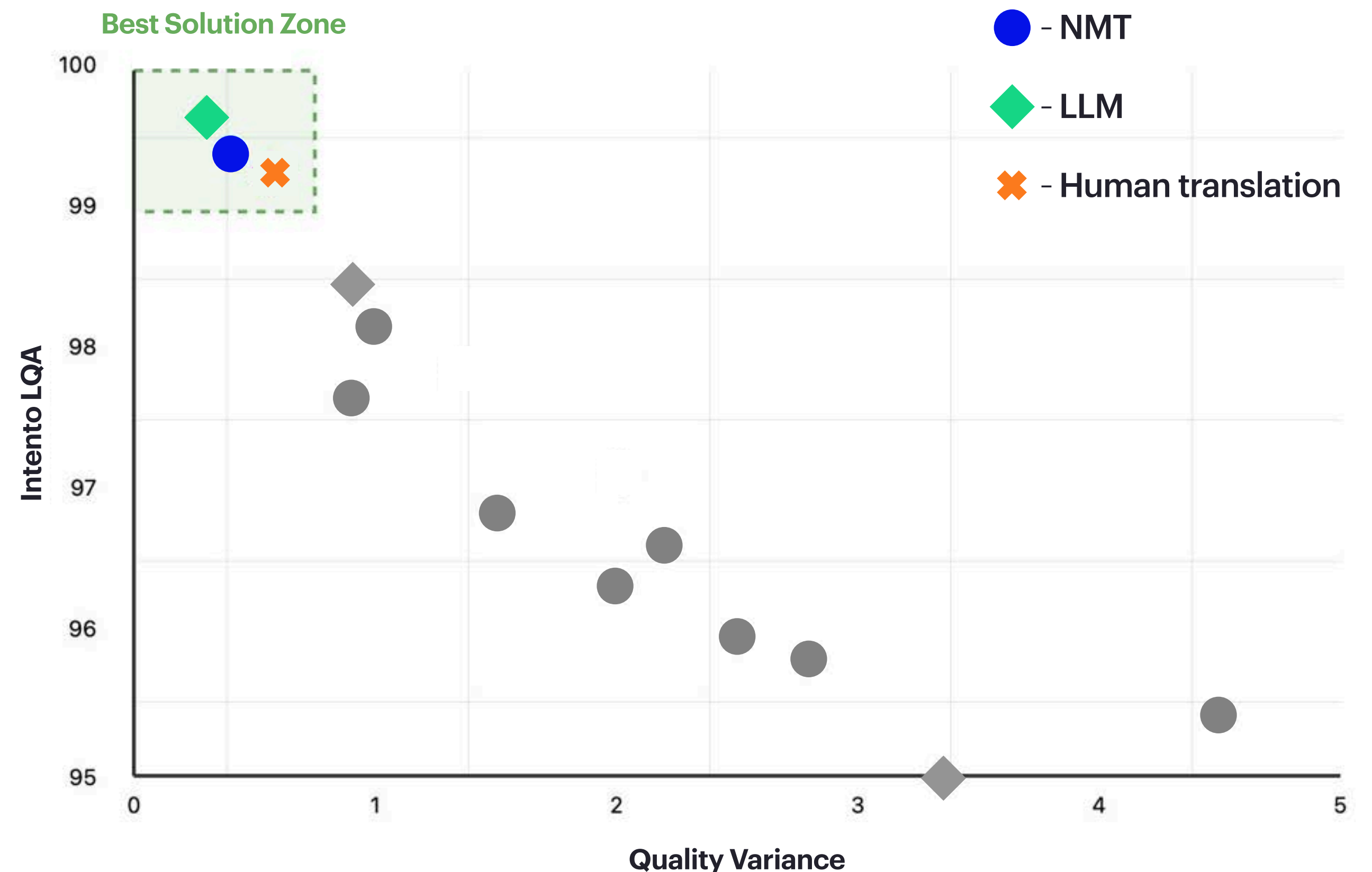
Intento LQA:

- Major error (5-point deduction) - Addition: The machine translation adds '欧州' (European) which is not present in the source text.
- Major error (5-point deduction) - Mistranslation: The machine translation uses 'つながった' (led to) which implies a past action, whereas the source text implies a future consequence.

3.7 How we Determine Best Solutions

- For every provider we build two 83% confidence interval¹ bands – one around the mean Intento LQA and one around the quality variation.
- Confidence intervals show the range where we expect a provider's true quality score to fall 83% of the time. This helps us avoid making decisions based on lucky or unlucky test results, ensuring we identify genuinely better solutions.
- Variance measures how consistent a provider's quality is across different translations - low variance means reliable, predictable quality while high variance indicates unpredictable results. Solutions with lower variance are more trustworthy for business use, as they deliver consistent quality.
- A provider is only considered to be “best” if both its quality and variance intervals overlap with the highest scoring provider.

1. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance?. J Insect Sci. 2003;3:34. doi:10.1093/jis/3.1.34.



3.8 How Do We Choose Solutions for Human Evaluation?

- We select 9 solutions for MQM-DQF-based Human LQA:
 - **Top-4 Models:** We select the four highest-performing models based on Intento LQA scores, regardless of whether they are an NMT engine or an LLM.
 - **Human Translation:** in this study, human translation is used not for the evaluation (“gold standard”), but as a reference point for the human translation quality (which as we know is not “perfect”).
 - **Multi-Agent Solution:** A multi-agent solution that combines translation, review and post-editing agents.
 - **Fair NMT Representation:** To counter potential LLM evaluation bias and ensure low-latency options, we include the top 3 NMT solutions minimum for each language pair.
- This approach provides good representation of LLMs and specialized NMT models, while keeping real-world needs in mind, as most LLM-based solutions can't handle real-time translation.
- By ensuring we test at least three NMT models per language pair, we account for a subtle risk: LLM-based evaluation metrics might naturally favor LLM-based translation solutions, potentially overlooking the strengths of traditional translation engines.

4. Automatic Quality Evaluation

- 4.1 Best Solutions by Automated Evaluation
- 4.2 Intento LQA vs Quality Variance Charts
- 4.3 Large Language Models Outperform Traditional NMT Solutions in Most Languages
- 4.4 Legal and Technical Content Exposes Issues off Human Translation
- 4.5 GPT-4.1 and Gemini 2.5 Pro Consistently Outperform Other Models
- 4.6 Requirements-Based Customization Benefits Vary for Low-Tier Engines
- 4.7 Mid-Tier Models Show Consistent Customization Benefits
- 4.8 High-Performance Models Achieve Noticeable Gains With Customization

4.1 Best Solutions by Automated Evaluation

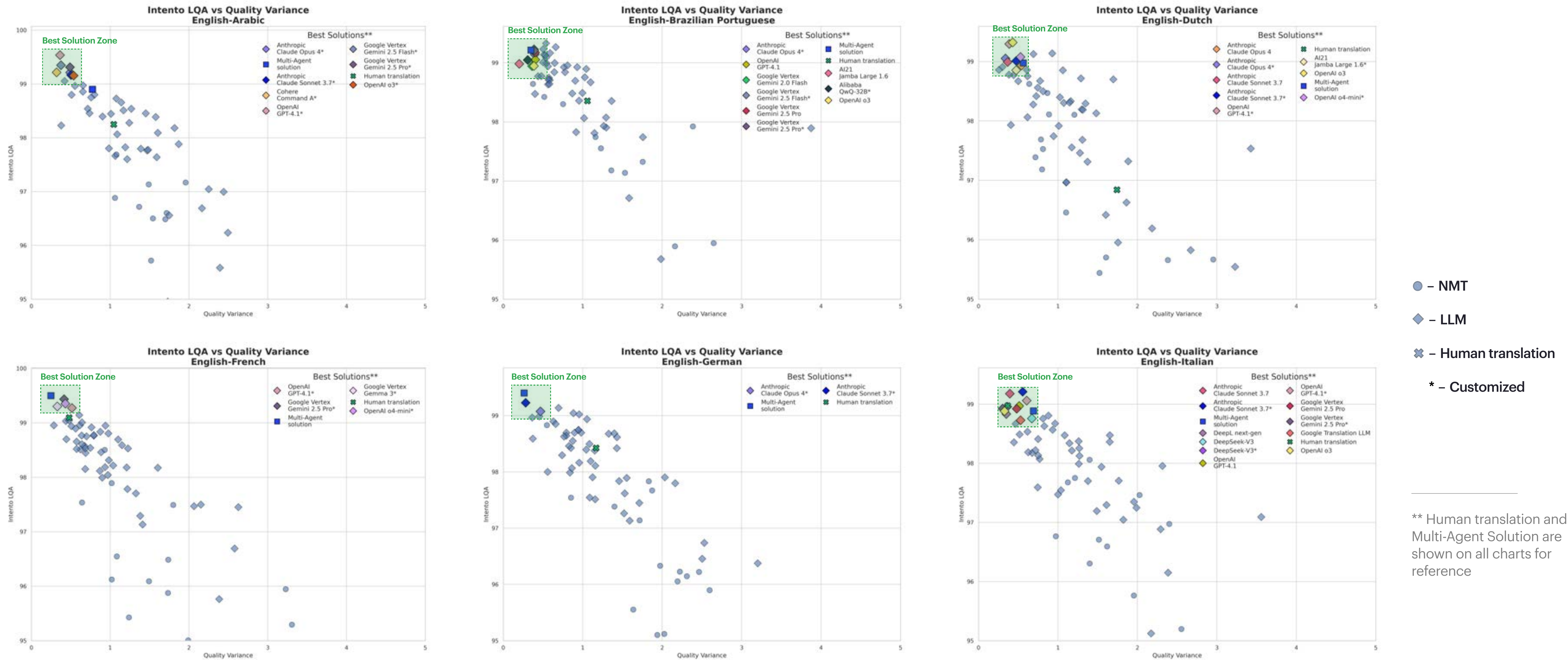
Best solutions by meeting translation requirements

	AI21	Alibaba	Anthropic	Cohere	DeepL	DeepSeek	Google	Translated	Meta	Multi-Agent	OpenAI
en-ar	Claude Opus 4	Claude Sonnet 3.7	Cohere Command A	Gemini 2.5 Flash	Gemini 2.5 Pro	GPT-4.1	OpenAI o3				
en-de	Claude Opus 4	Claude Sonnet 3.7	Multi-Agent Solution								
en-es	Claude Sonnet 4	Gemini 2.5 Flash	GPT-4.1	Llama 4 Maverick	Multi-Agent Solution	OpenAI o4-mini					
en-fr	Gemini 2.5 Pro	Google Gemma 3	GPT-4.1	Multi-Agent Solution	OpenAI o4-mini						
en-it	Claude Sonnet 3.7	DeepL next-gen	DeepSeek-V3	Gemini 2.5 Pro	Google Translation LLM	GPT-4.1	Multi-Agent Solution	OpenAI o3			
en-ja	Claude Opus 4	DeepSeek-V3									
en-ko	Claude Sonnet 3.7	Claude Sonnet 4	Gemini 2.0 Flash	DeepSeek-V3	Gemini 2.5 Flash	Gemini 2.5 Pro	GPT-4.1	OpenAI o4-mini	OpenAI o3		
en-nl	AI21 Jamba Large 1.6	Claude Opus 4	Claude Sonnet 3.7	GPT-4.1	Multi-Agent Solution	OpenAI o3	OpenAI o4-mini				
en-pt	AI21 Jamba Large 1.6	Alibaba QwQ-32B	Claude Opus 4	Gemini 2.0 Flash	Gemini 2.5 Flash	Gemini 2.5 Pro	GPT-4.1	Multi-Agent Solution	OpenAI o3		
en-uk	Gemini 2.5 Flash	Gemini 2.5 Pro	GPT-4.1	Llama 4 Maverick	Multi-Agent Solution	OpenAI o3	Translated Lara				
en-zh	Gemini 2.5 Pro	OpenAI o3	OpenAI o4-mini								

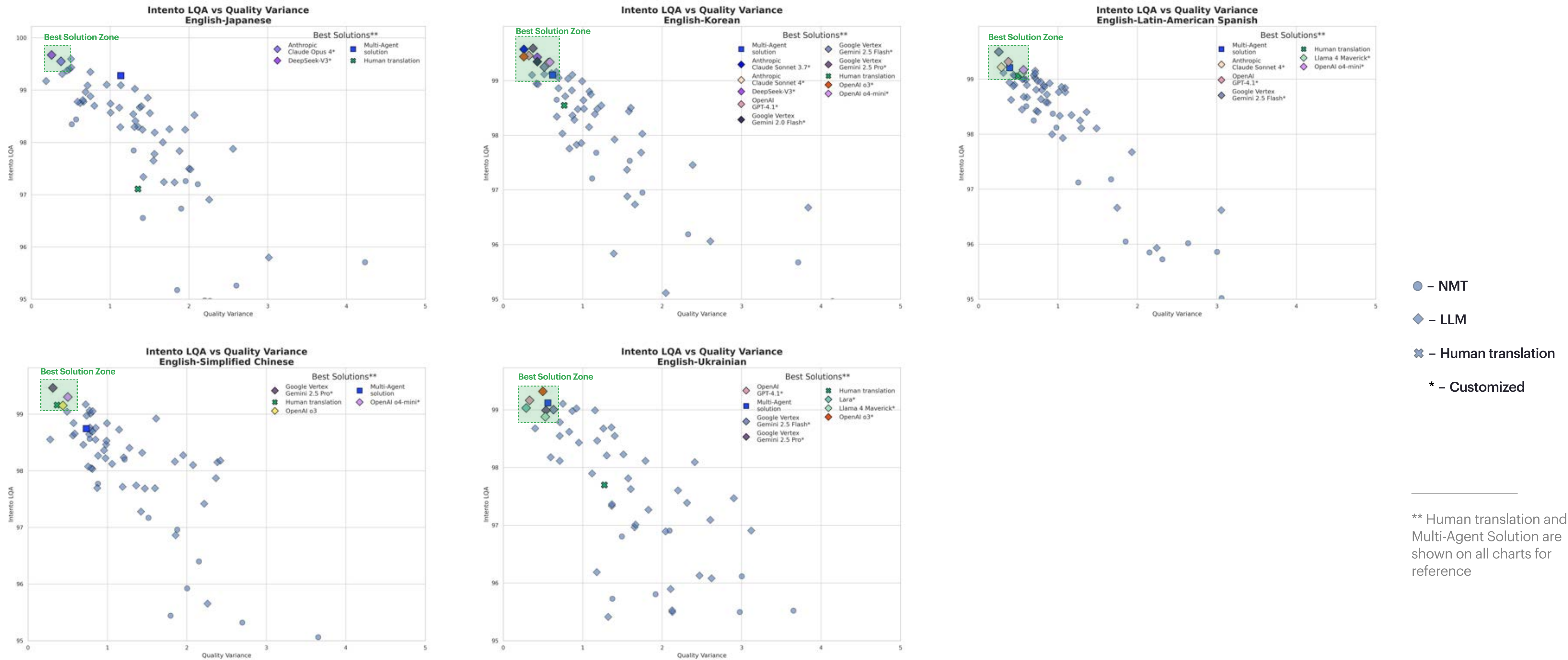
– Customized – Multi-Agent – Baseline

* A solution is only considered to be “best” if both its quality and variance intervals overlap with the highest scoring solution (see Section 3.7).

4.2 Intento LQA vs Quality Variance Charts (1/2)



4.2 Intento LQA vs Quality Variance Charts (2/2)



4.3 Large Language Models Outperform Traditional NMT Solutions in Most Languages

- [Anthropic's Claude](#), [OpenAI's GPT-4.1](#) and [Google Vertex' Gemini 2.5 Pro](#) models consistently appear as top performers across all languages, both as the baseline and customized solutions.
- [Multi-Agent](#) solution achieves high scores and great translation consistency in most languages, shining in [Brazilian Portuguese](#), [French](#), [German](#) and [Japanese](#) in particular.
- [Arabic](#) showcases a larger variation in quality than most other languages, due to heavy syntactic struggles with full text translation for a noticeable number of solutions.
- [Romance and Germanic languages](#) ([French](#), [German](#), [Dutch](#), [Italian](#), [Spanish](#), and [Brazilian Portuguese](#)) show less variation between solutions due to high-resource nature.
- [Asian languages](#) ([Japanese](#), [Korean](#), [Chinese](#)) display much wider spreads in both quality and deviation, indicating these are more challenging with clearer separation between good and poor performers.
- Large Language Models fine-tuned for translation, such as [Google Translation LLM](#), [DeepL next-gen](#) and [Lara](#), achieve great results in [Italian](#) and [Ukrainian](#).
- [Ukrainian](#) demonstrates the most variance in solutions' quality among all languages, mostly due to being lower-resources and harder in grammar and syntax than most languages.

4.4 Legal and Technical Content Exposes Issues off Human Translation

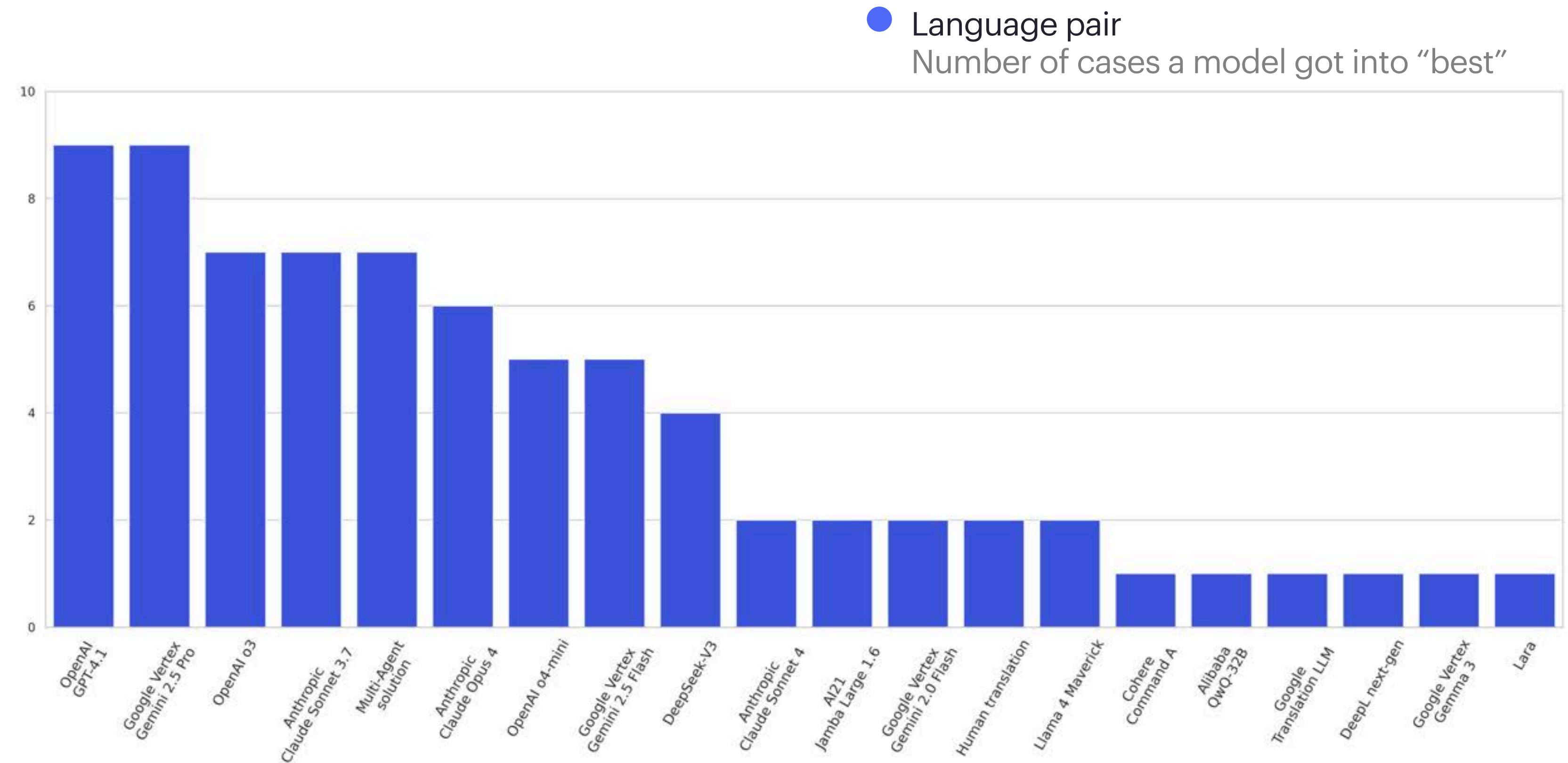
- **Arabic** human translation suffers from translations that read as stilted and unnatural, and inconsistencies in terminology use.
- **Brazilian Portuguese** translations showcase some mistranslations, with such examples as "misrepresentation" morphing into "distortion" and other key concepts losing their precision.
- In **Dutch**, most issues stem from legal language, with critical data protection concepts getting lost in translation or disappearing entirely through omissions.
- Major issues in **German** human translations include mistranslations and tone of voice issues.
- **Korean** translations suffer from switching between semi-formal and formal tone throughout the same text, and unnecessary meaning changes.
- In **Japanese**, there are multiple issues with opposite meaning and markup, resulting in far lower scores than automatic solutions.
- **Ukrainian** human translations have some factual mistakes and major under-translations, simplifying legal and technical concepts where not necessarily required.
- **NOTE:** In this study, human translation is used not for the evaluation ("gold standard"), but as a reference point for the human translation quality (which as we know is not "perfect").

4.5 GPT-4.1 and Gemini 2.5 Pro Consistently Outperform Other Models

11 language pairs

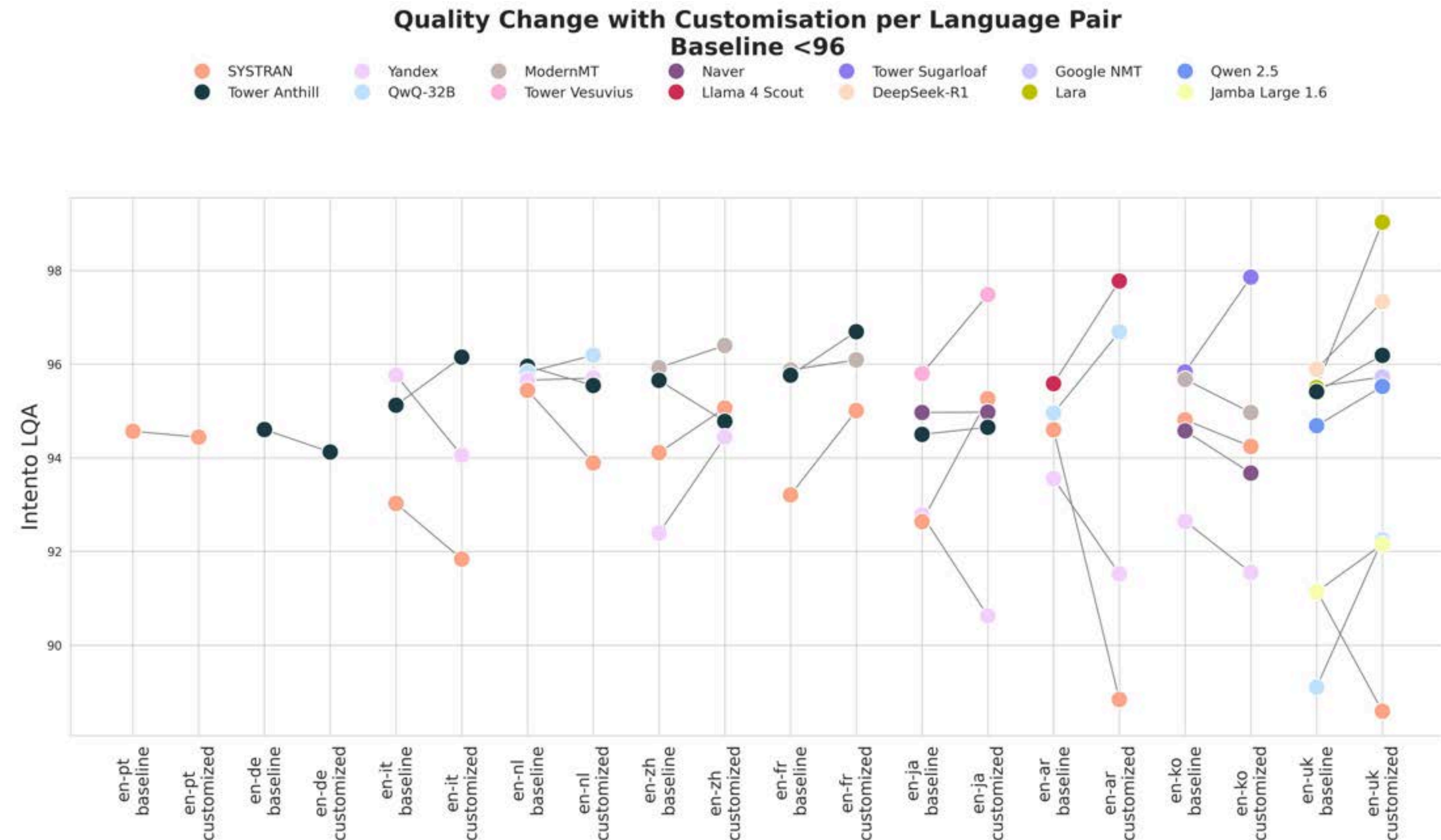
Some models were tested only in their specific language pairs:

- HiThink RoyalFlush specializes in en-zh translation
- TREBE specializes in Iberian languages, and was used for en-es translation
- Tarjama specialized in Arabic translation



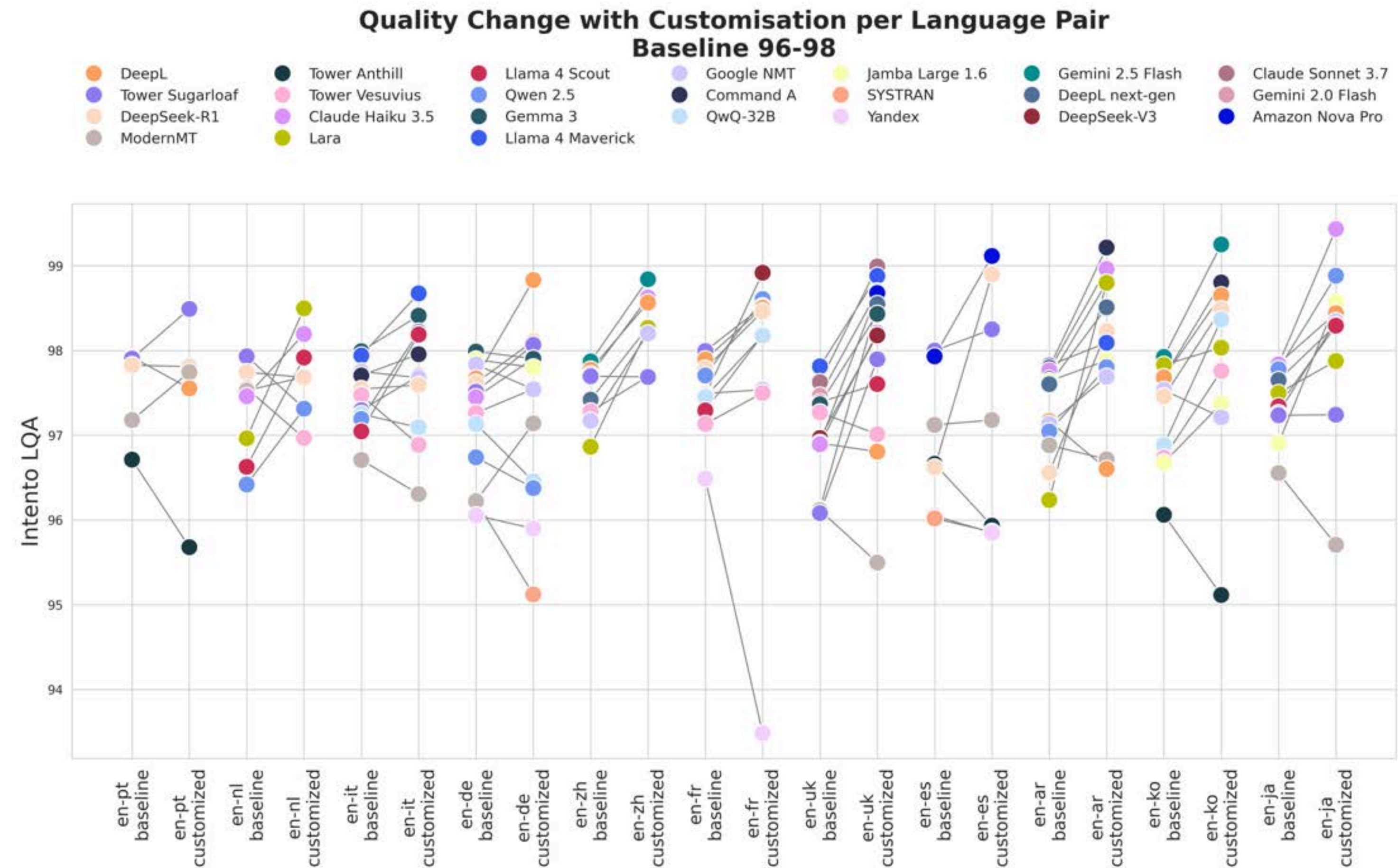
4.6 Requirements-Based Customization Benefits Vary for Low-Tier Engines

- Engines starting with a baseline score of Intento LQA below 96 show highly variable improvements.
- Translation-tuned LLMs, such as [Tower Anthill](#) and [Lara](#), consistently improve from glossary-based customization.
- [SYSTRAN](#) and [Yandex](#) suffer more than benefit from glossary addition in most language pairs because of tag disruption and grammatically incorrect terminology use that degrades the quality; noticeably, in Simplified Chinese, without morphology, the quality improves with glossary addition for both models.



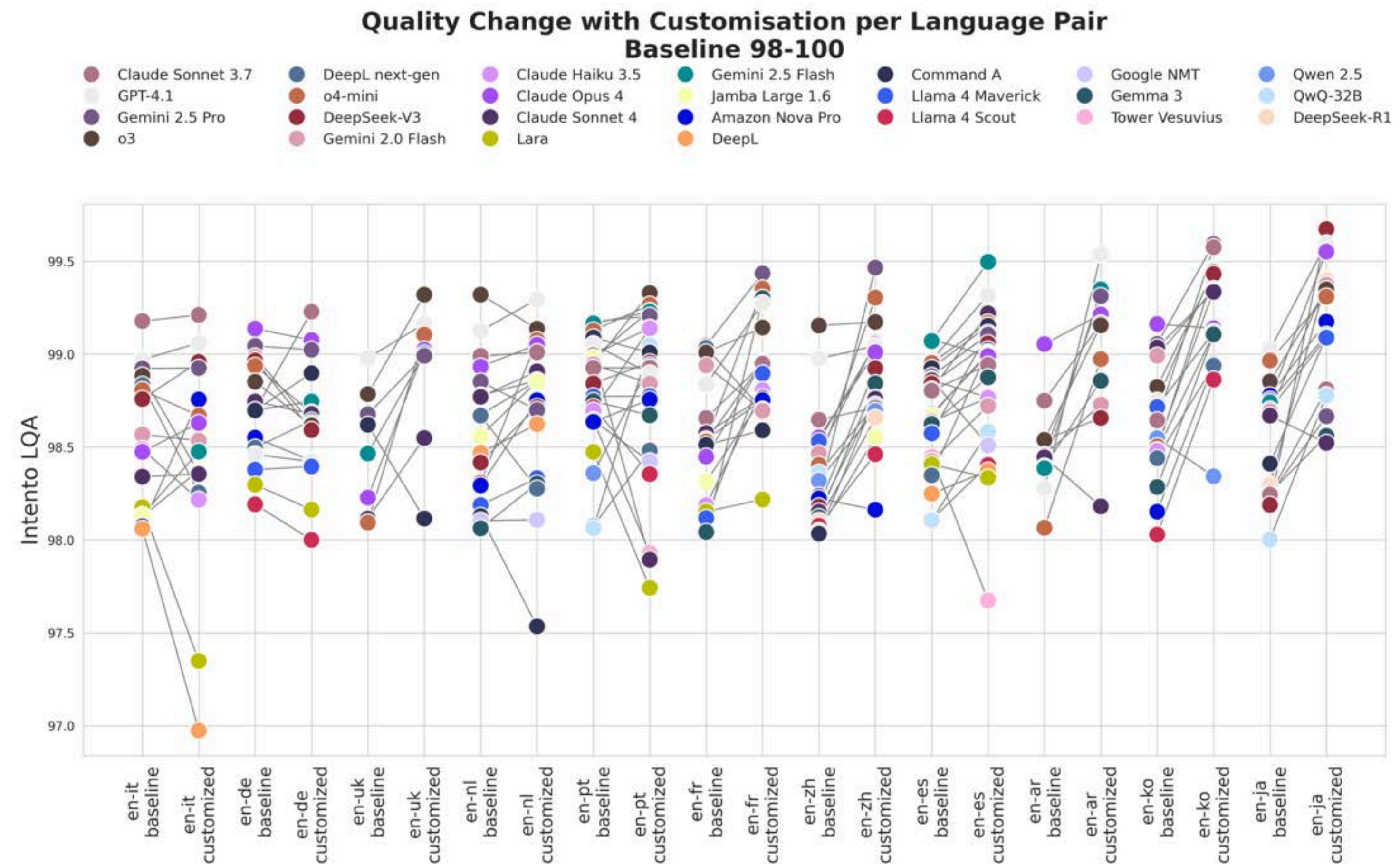
4.7 Mid-Tier Models Show Consistent Customization Benefits

- [DeepL](#) and [DeepL next-gen](#) consistently improve across language pairs.
- The majority of models in this category achieve notable improvements on the Intento LQA scale, with some standout performers like [Command A](#) and [Gemini 2.5 Flash](#) reaching scores above 99 after customization.
- [Yandex](#) achieves higher scores in European languages, such as French, in the baseline solution but still sees some tag usage degradation with glossary introduction.
- Generally, a rare drop in LLM translation quality with the addition of requirements comes from incorrect instruction following (i.e. re-writing the translation more than tone and glossaries require).



4.8 High-Performance Models Achieve Noticeable Gains With Customization

- Translation models starting with baseline scores between 98-100 on the Intento LQA scale show the most consistent improvements from customization.
- Several Large Language Models, including [Claude Opus 4](#), [Claude Sonnet](#) variants, [GPT-4.1](#), and [o4-mini](#) all demonstrate reliable quality improvements without any significant degradation risks.
- Drops in quality for [Lara](#) and [DeepL](#) in Italian are easily explainable: Lara mixes formal and informal registries together in several texts despite instructions, while [DeepL](#) produces more literal translations and is less contextually aware when glossary is enforced.
- [Korean](#) and [Japanese](#) show exceptional improvements, suggesting that customization remains valuable for challenging languages.

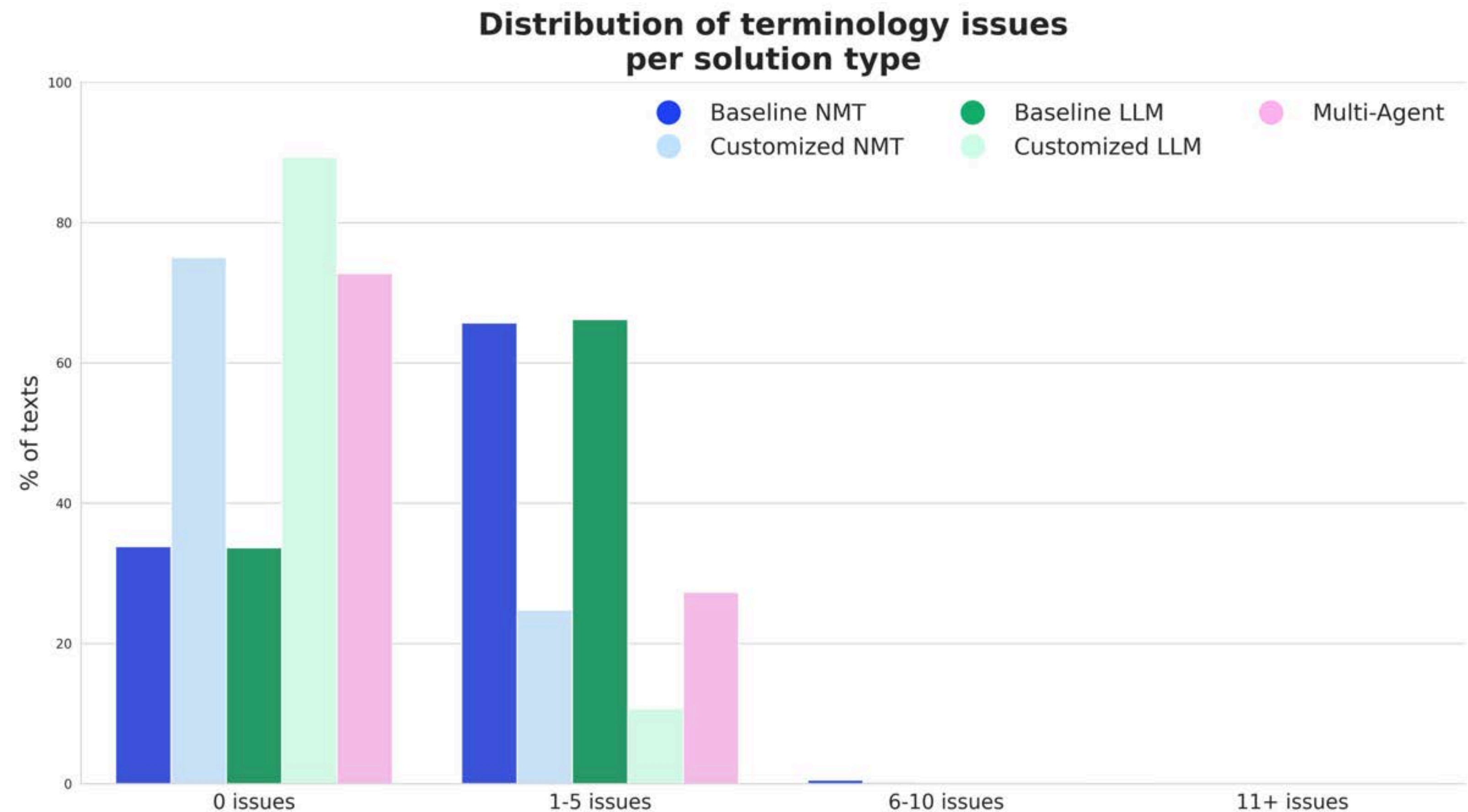


5. Automatic Requirements-Based Analysis

- 5.1 Terminology Accuracy Demands Customized Solutions
- 5.2 Top Solutions Excel at Terminology Consistency With Customization
- 5.3 Tone-of-Voice Compliance Varies Significantly Across Solution Types
- 5.4 Customization Yields Consistent Improvements in Tone Consistency
- 5.5 Multi-Agent Workflow Excels at Translation Consistency
- 5.6 Solution Response to Customization Varies Widely
- 5.7 All Solutions Achieve Near Perfection In Tag Handling
- 5.6 Arabic RTL Remains the Final Frontier of Tag Processing

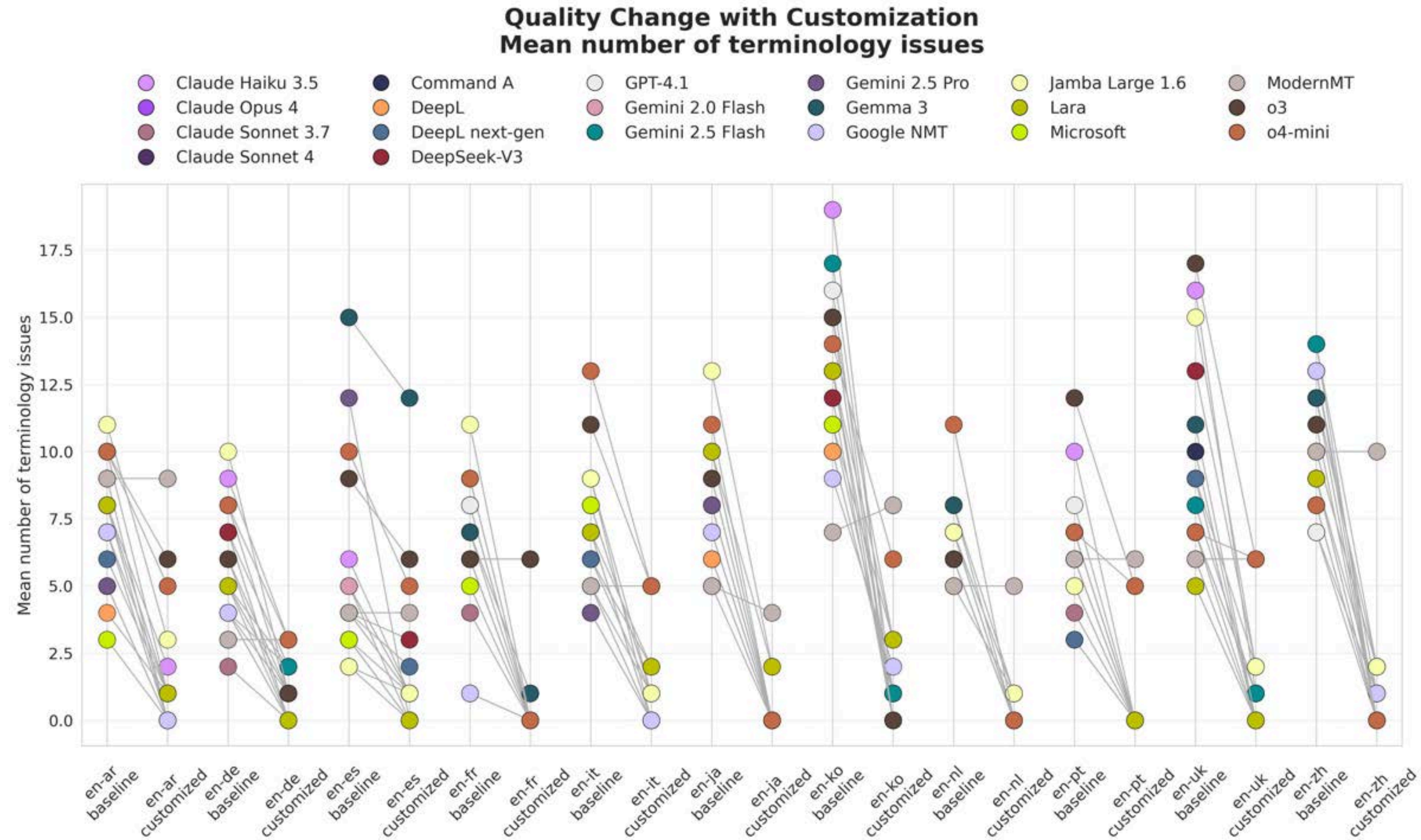
5.1 Terminology Accuracy Demands Customized Solutions

- LLMs customized with terminology injection produce the least terminology errors, with customized NMTs and Multi-Agent solution following closely.
- Both baseline LLMs and NMT engines typically produce 1-5 terminology issues per text.
- Multi-Agent solution relies heavily on the Reviewer Agent—if it misses errors from the Translation Agent, subsequent steps will fail. Custom Reviewer Agents tailored to specific needs are therefore crucial.
- Content with regulated, safety-critical, or branded terms requires terminology-aware solutions (customized LLM or a Multi-Agent solution).



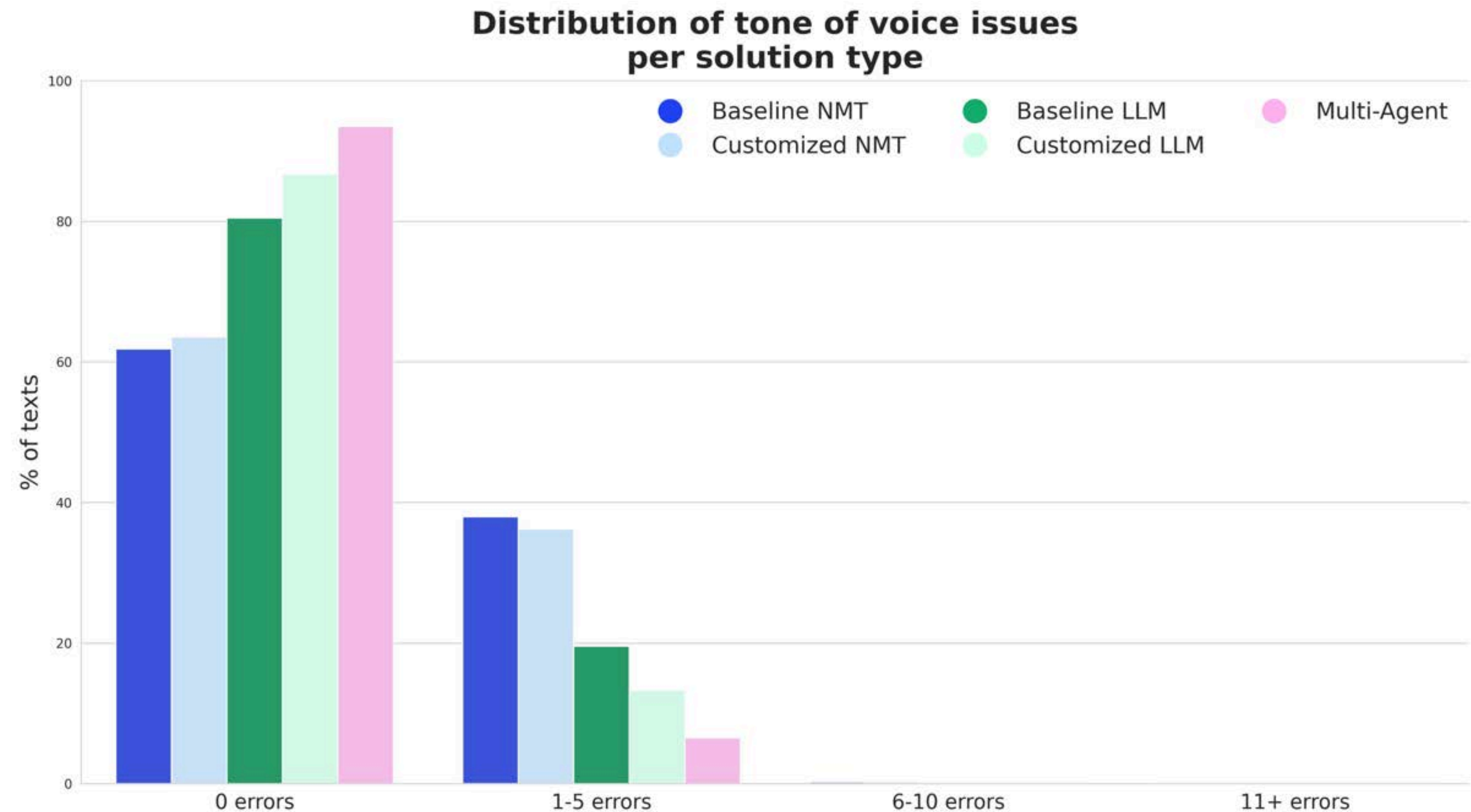
5.2 Top Solutions Excel at Terminology Consistency With Customization

- NMT engines with glossary support and LLMs with terminology-specific instructions show immense improvement in terminology handling, having less than 1 inconsistently translated term per text.
- LLMs that lack domain-specific terminology, such as [o3](#), [Jamba Large 1.6](#), and [Claude Haiku 3.5](#), reduce errors by roughly one-half after terminology injection via RAG.
- [ModernMT](#) does not handle Korean terminology well, while [Gemma 3](#) does not get as much improvement as other solutions in English>Spanish.



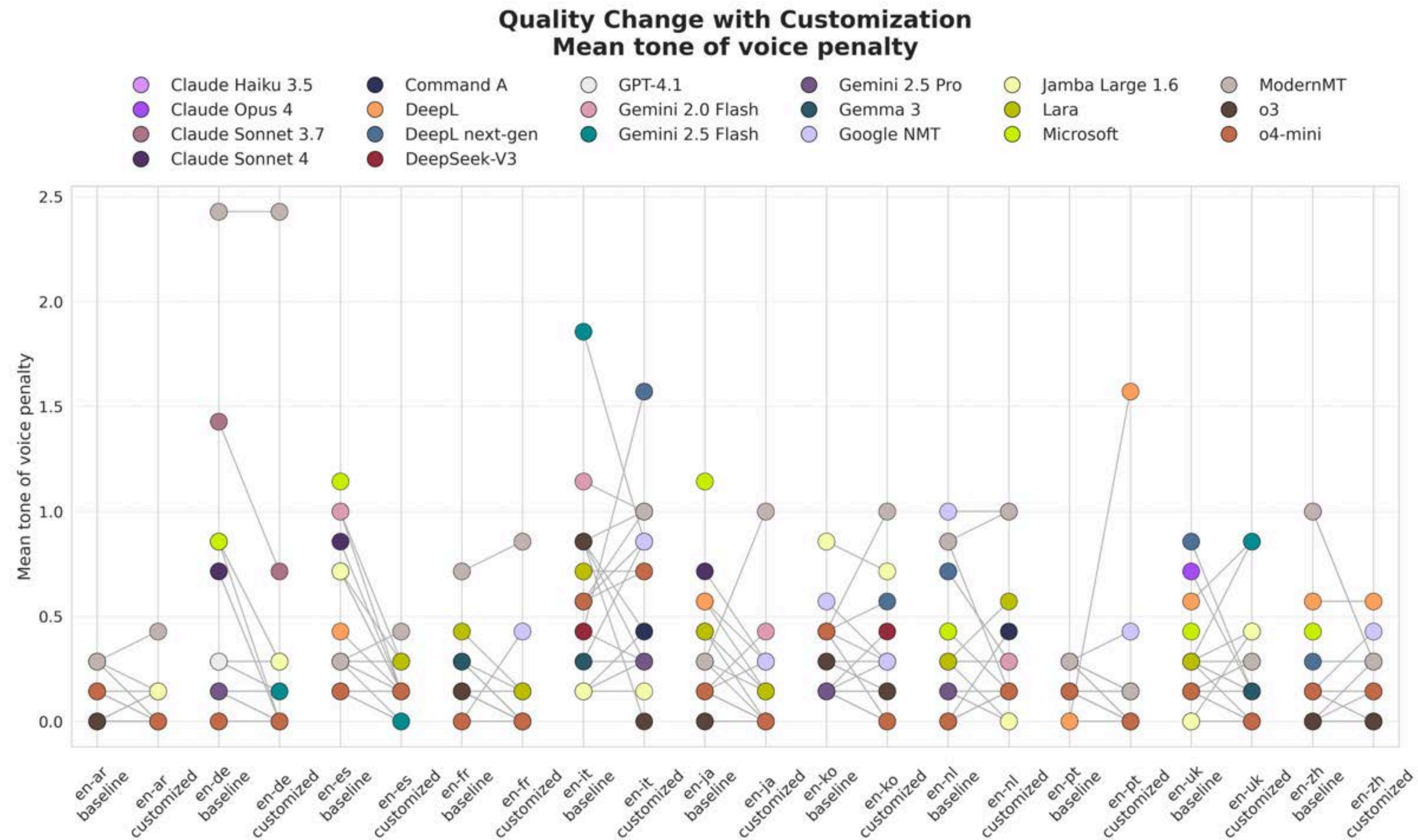
5.3 Tone-of-Voice Compliance Varies Significantly Across Solution Types

- [Multi-Agent](#) solution and [LLMs customized with tone-of-voice injections](#) achieve near ideal compliance with tone-of-voice requirements.
- [Baseline LLMs](#) perform reasonably well, with approximately 50% less formal tone consistency problems compared to [baseline NMT engines](#). However, 15-20% of LLMs still produce at least one tone error in their translations.
- For content where tone is critical, for example, customer-facing or regulated materials, we recommend using LLMs with specific tone of voice instructions rather than baseline machine translation.



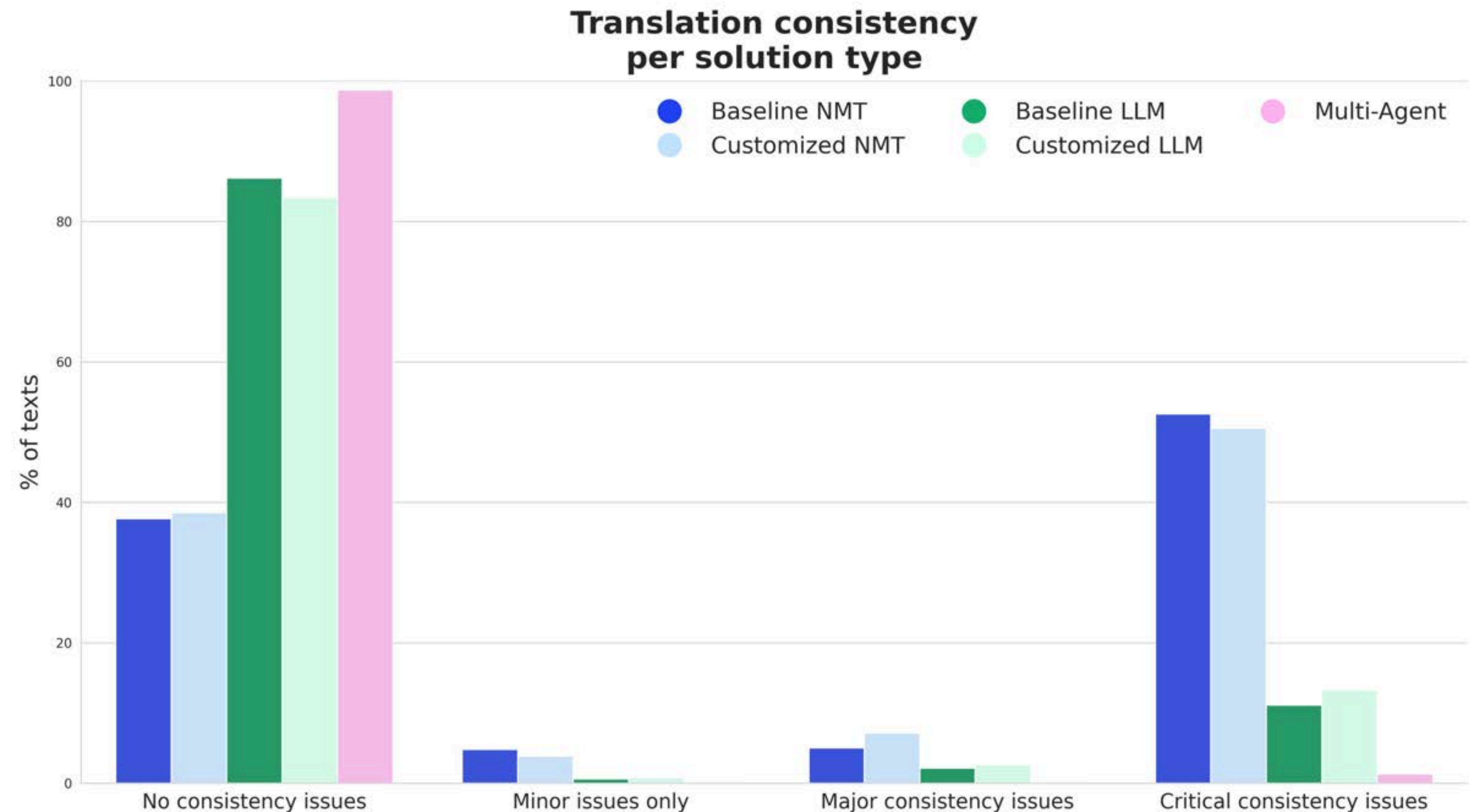
5.4 Customization Yields Consistent Improvements in Tone Consistency

- NMTs with adjustable tone of voice and LLMs noticeably improve tone of voice management in languages with clear formality patterns.
- Only a few engines degrade in terms of tone requirements when customization is involved due to other customization aspects (i.e. terminology support) affecting the overall style and producing some inconsistencies.
- Overall, most NMT engines with formality support and LLMs with tone-specific instructions quickly adapt to the desired tone.



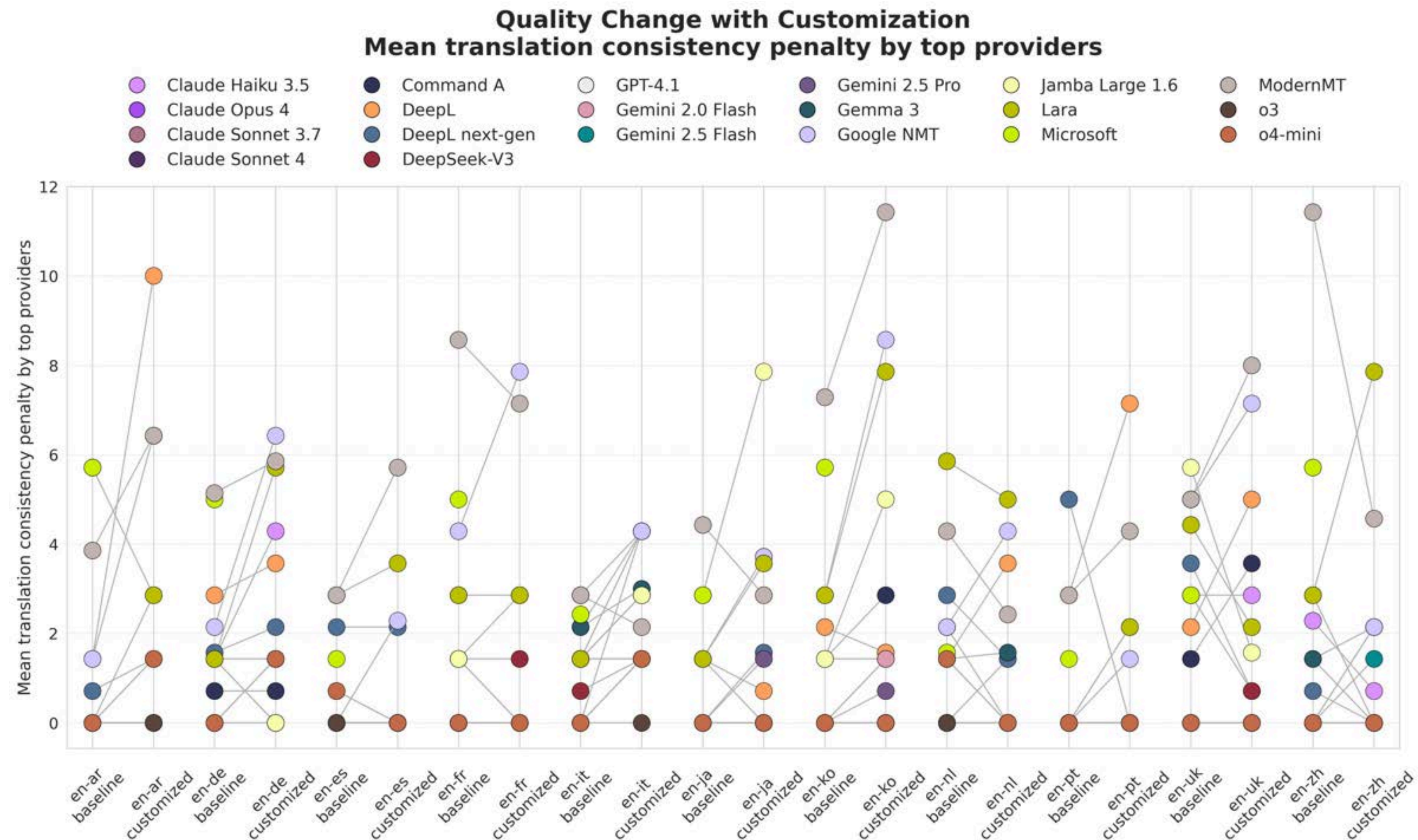
5.5 Multi-Agent Workflow Excels at Translation Consistency

- **Multi-Agent** solution provides the most dependable output in terms of referential consistency across the text, stylistic and structural consistency.
- Large context window allows both **baseline** and **customized LLMs** to achieve consistent translation through the text in 90% of cases.
- **NMT** engines can rarely achieve truly consistent output due to most engines splitting the text into segments or fragments as they have been trained this way.



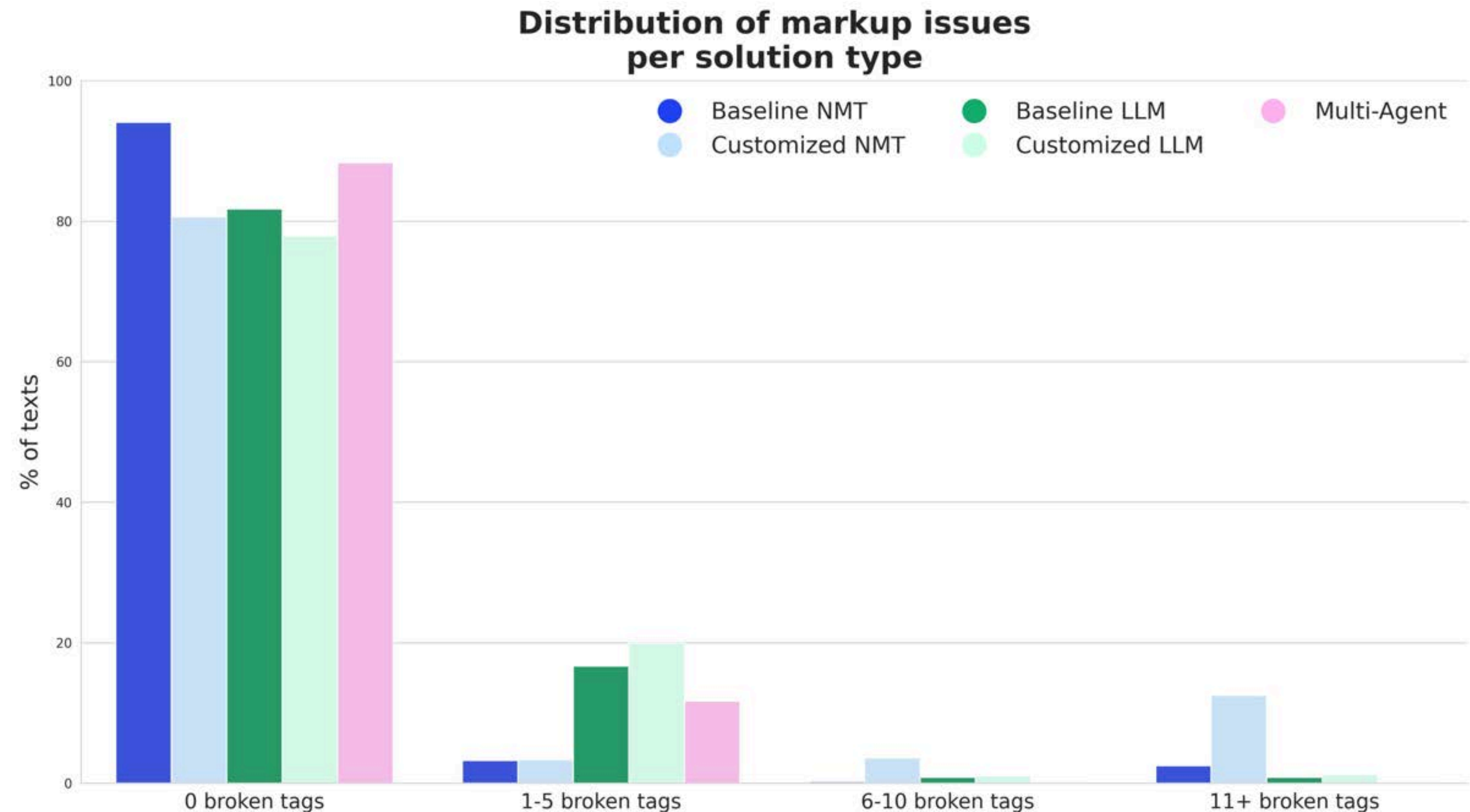
5.6 Solution Response to Customization Varies Widely

- Translation consistency remains the most challenging requirement across all solutions types, with penalties ranging from near-zero to over 10 points depending on the solution-language combination.
- Japanese and Korean translations particularly benefit from customization across most models.
- LLM-based solutions show more dynamic response to customization compared to traditional NMT engines, suggesting that LLMs' large context windows help achieve better referential and stylistic consistency.



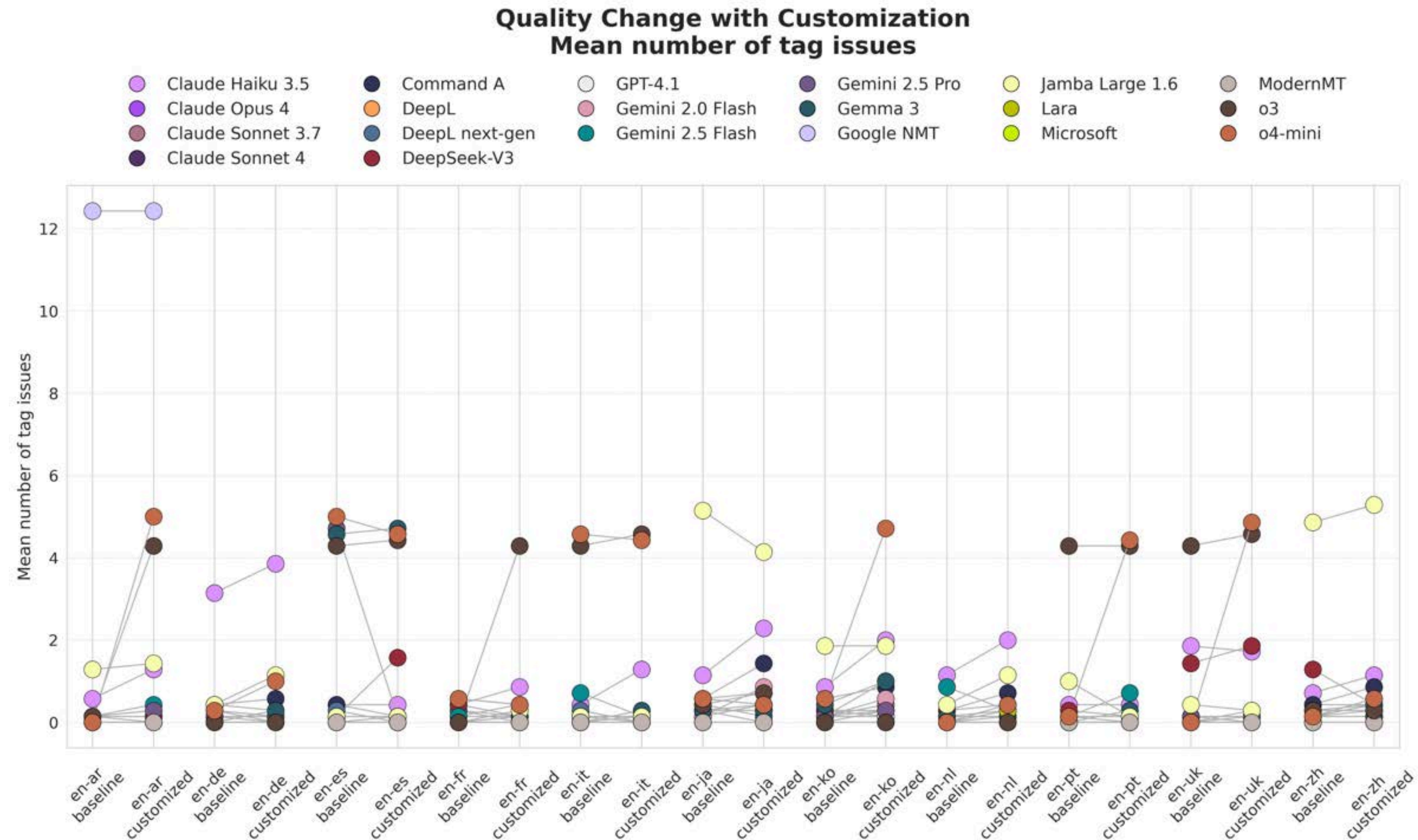
5.7 All Solutions Achieve Near Perfection In Tag Handling

- There is very notable progress in how **NMT** engines treat HTML formatting compared to just a few years ago, as **baseline** solutions achieve nearly ideal tag processing.
- **Customized NMT** notably suffers when it comes to formatting requirements as additional of terminology breaks sentence and tag structure of several models.
- **LLM-based** solutions, including **Multi-Agent**, have little to no trouble when it comes to formatting requirements.



5.8 Arabic RTL Remains the Final Frontier of Tag Processing

- The majority of solutions cluster near zero issues even before customization, reflecting industry-wide improvements in HTML and XML tag processing over recent years.
- Arabic very notably is the hardest language for tag handling due to the difference of writing direction, right to left, which can introduce difficulties in text formatting for some models.
- For some LLMs, such as [Claude Haiku 3.5](#), addition of glossary and style-specific instructions can lead to some tag handling inconsistencies.



6. Human LQA Analysis

- 6.1 Top Providers Evaluated in Human LQA
- 6.2 Human LQA Adjustment to Report Requirements
- 6.3 Human LQA vs Quality Variance Charts
- 6.4 The Human-AI Translation Gap Has Virtually Disappeared
- 6.5 Error Analysis Reveals Language-Specific Translation Challenges
- 6.6 Human LQA Highlights Language Disparities
- 6.7 Ukrainian Stands as MT's Biggest Challenge While Chinese Excels
- 6.8 Best Solutions per Language Pair Based on Human LQA
- 6.9 Multi-Agent Solution Outperforms All Others by a Landslide
- 6.10 Several Google Solutions Perform On Par With the Multi-Agent Approach

6.1 Top Providers Evaluated in Human LQA

Top Providers Evaluated in Human LQA

	AI21	Anthropic	Cohere	DeepL	DeepSeek	Google	Microsoft	Multi-Agent	OpenAI	Translated					
en-ar			Cohere Command A	DeepL next-gen	Gemini 2.5 Flash	Gemini 2.5 Pro	Google NMT		GPT-4.1	Human translation	Multi-Agent Solution	Translated Lara			
en-de		Claude Opus 4		Claude Sonnet 3.7	DeepL	Gemini 2.0 Flash	Gemini 2.5 Pro	Google NMT		Human translation	Multi-Agent Solution	Translated Lara			
en-es		Claude Sonnet 4		DeepL next-gen	Gemini 2.5 Flash	Google NMT		GPT-4.1		Human translation	Multi-Agent Solution	OpenAI o4-mini	Translated Lara		
en-fr			DeepL next-gen		Gemini 2.5 Pro	Google Gemma 3	Google NMT		GPT-4.1		Human translation	Multi-Agent Solution	OpenAI o4-mini	Translated Lara	
en-it		Claude Sonnet 3.7		DeepL next-gen	DeepSeek-V3	Gemini 2.5 Pro	Google NMT		GPT-4.1		Human translation	Multi-Agent Solution		Translated Lara	
en-ja		Claude Haiku 3.5		Claude Opus 4	DeepL	DeepSeek-V3	Google NMT		GPT-4.1		Human translation	Multi-Agent Solution		Translated Lara	
en-ko		Claude Sonnet 3.7		DeepL next-gen		Gemini 2.5 Pro	Google NMT		GPT-4.1		Human translation	Multi-Agent Solution	OpenAI o3	Translated Lara	
en-nl		Claude Opus 4		DeepL next-gen		Google NMT		GPT-4.1		Human translation	Multi-Agent Solution	OpenAI o3	OpenAI o4-mini	Translated Lara	
en-pt	AI21	Jamba Large 1.6		DeepL next-gen		Gemini 2.5 Flash	Google NMT		Human translation		Multi-Agent Solution	OpenAI o3	OpenAI o4-mini	Translated Lara	
en-uk			DeepL next-gen		GPT-4.1		Human translation		Microsoft		ModernMT	Multi-Agent Solution	OpenAI o3	OpenAI o4-mini	Translated Lara
en-zh			DeepL		Gemini 2.0 Flash		Gemini 2.5 Pro	Google NMT		Human translation		Multi-Agent Solution	OpenAI o3	OpenAI o4-mini	Translated Lara

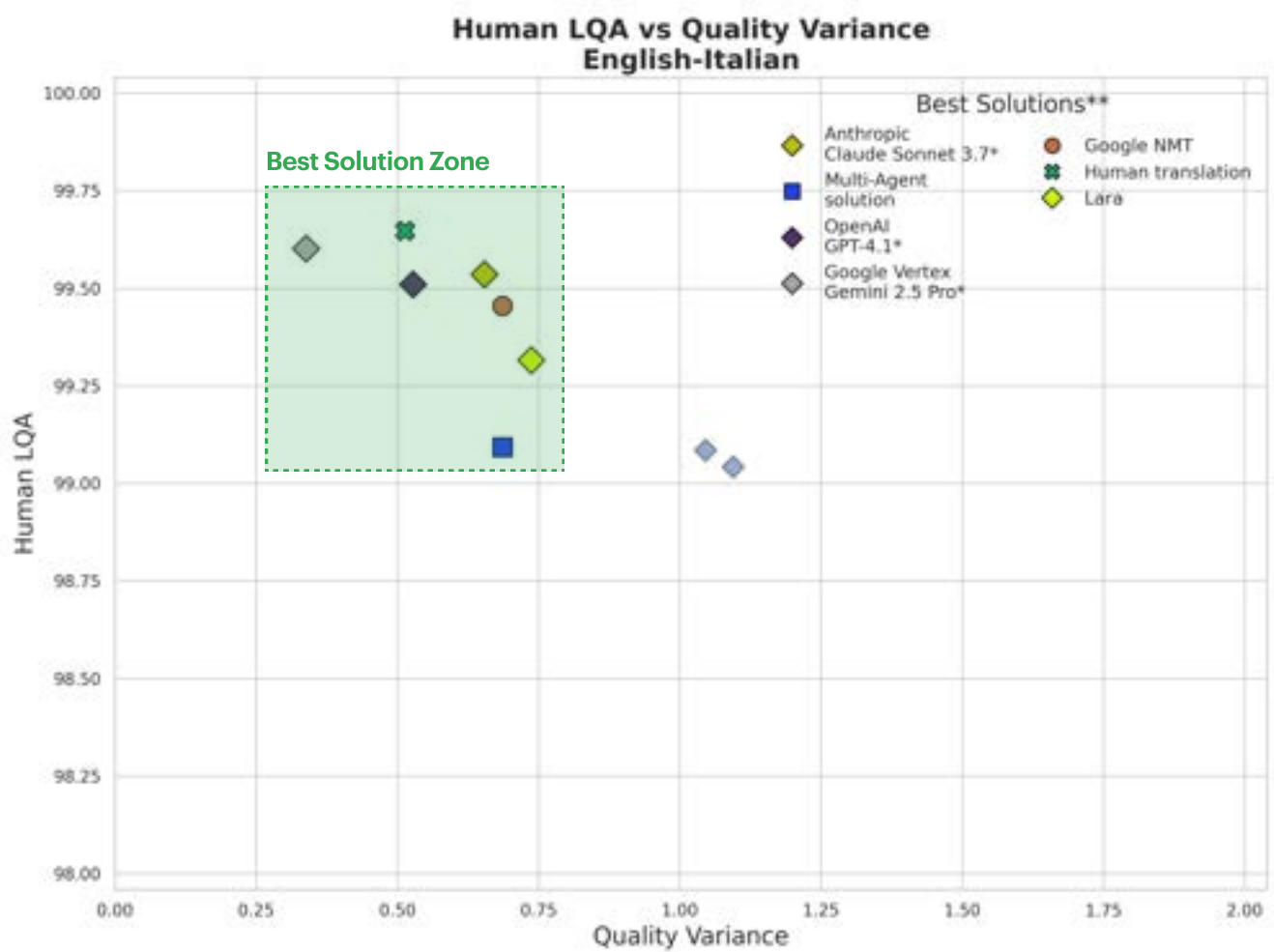
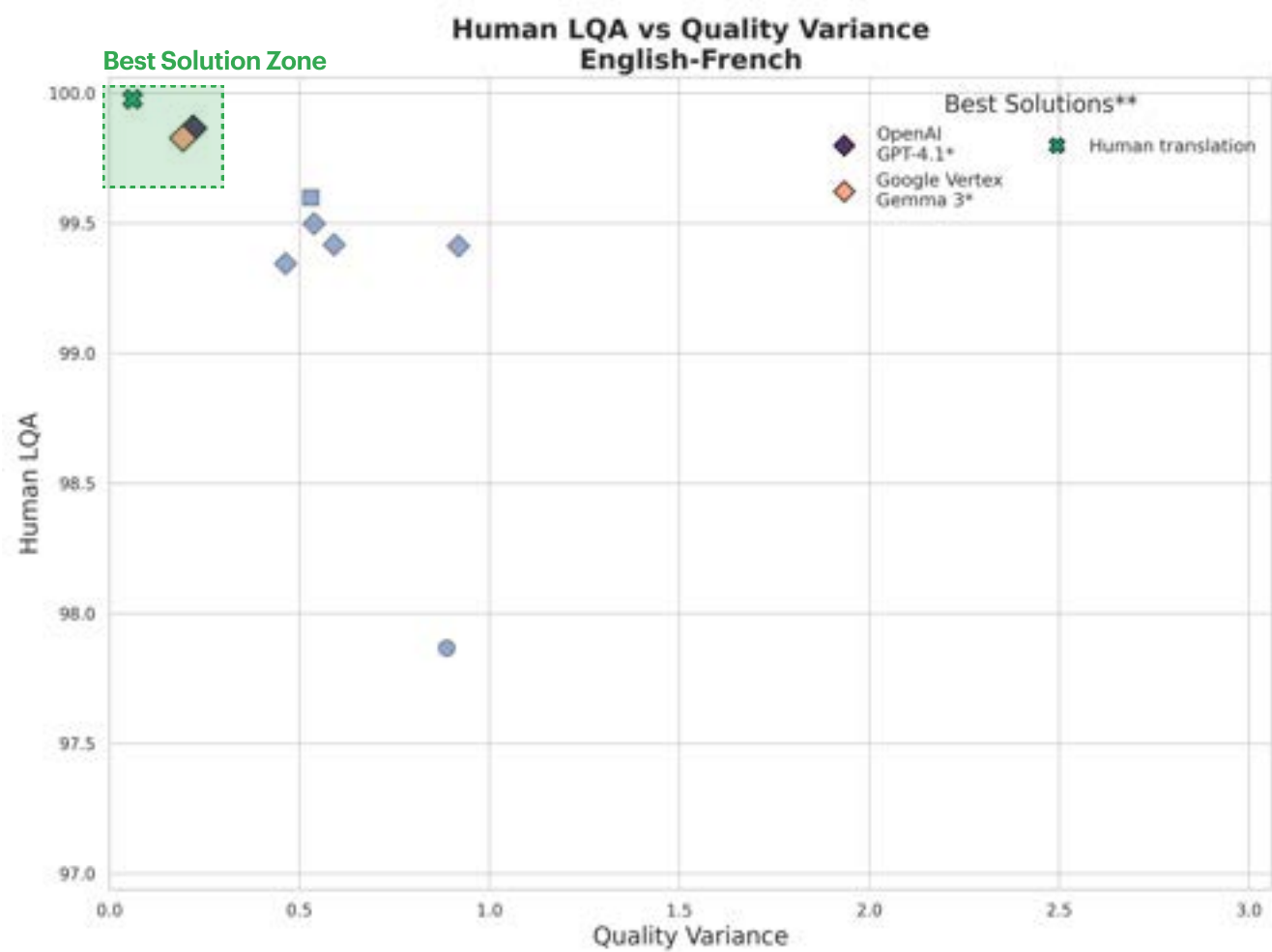
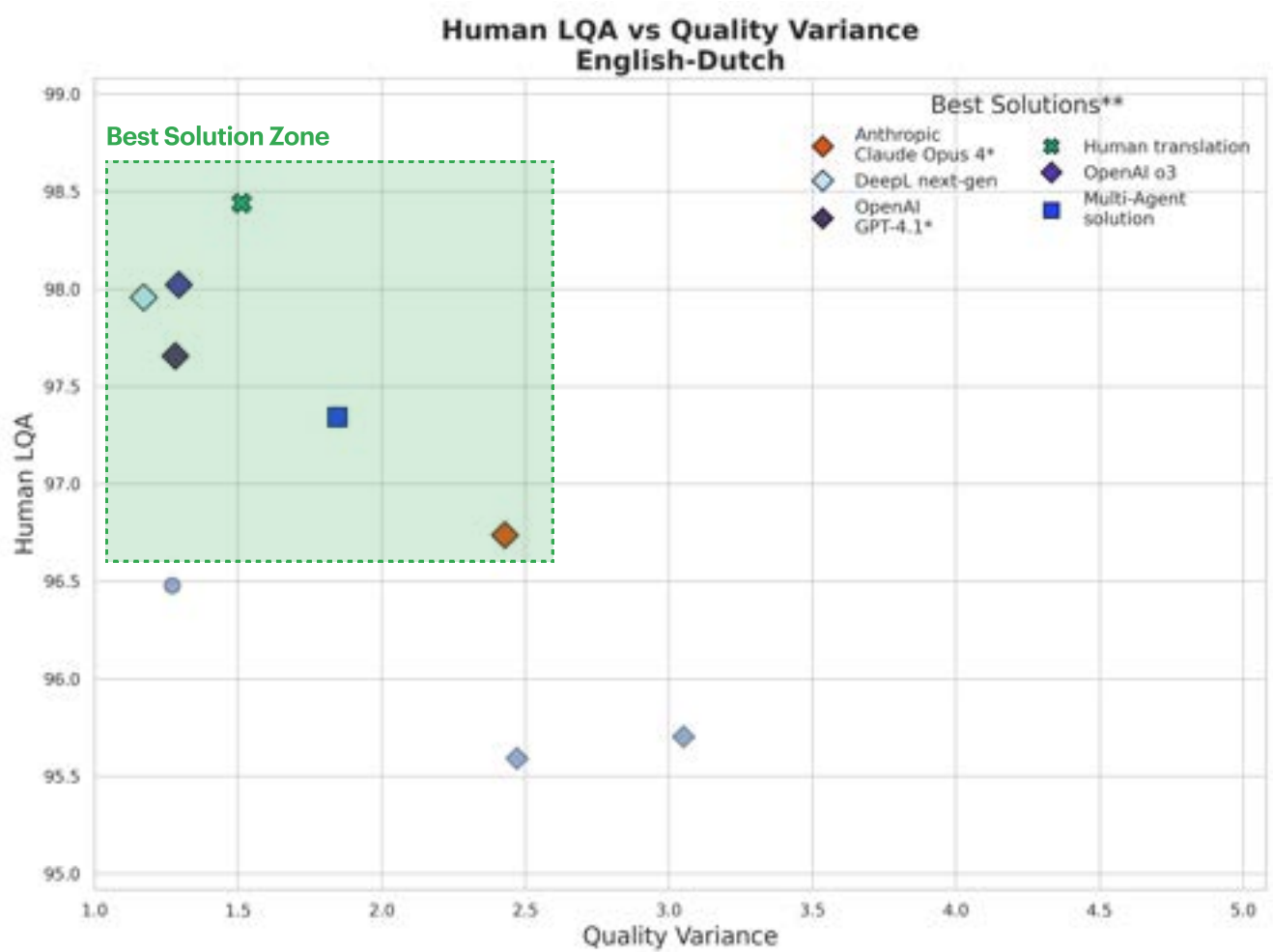
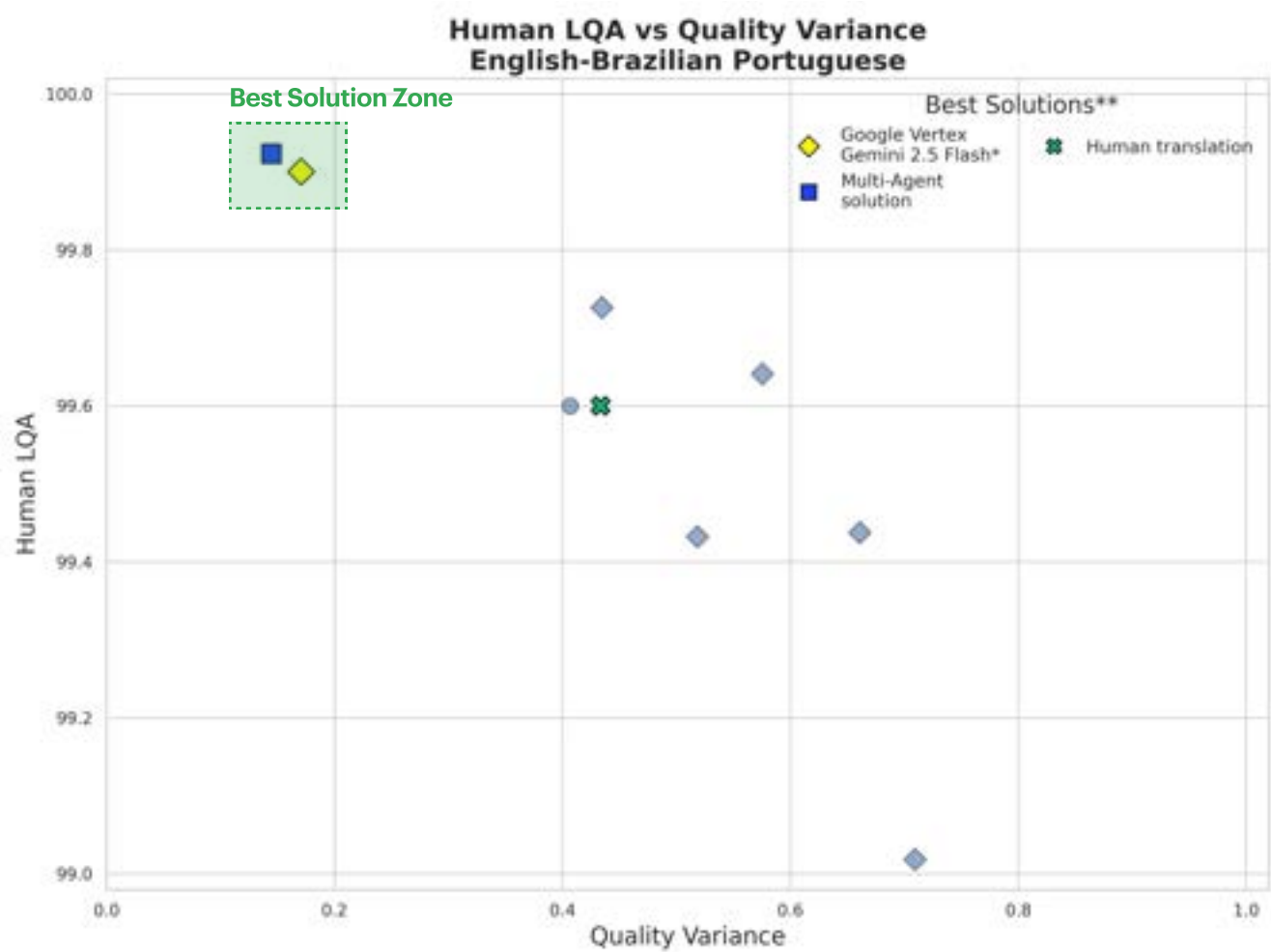
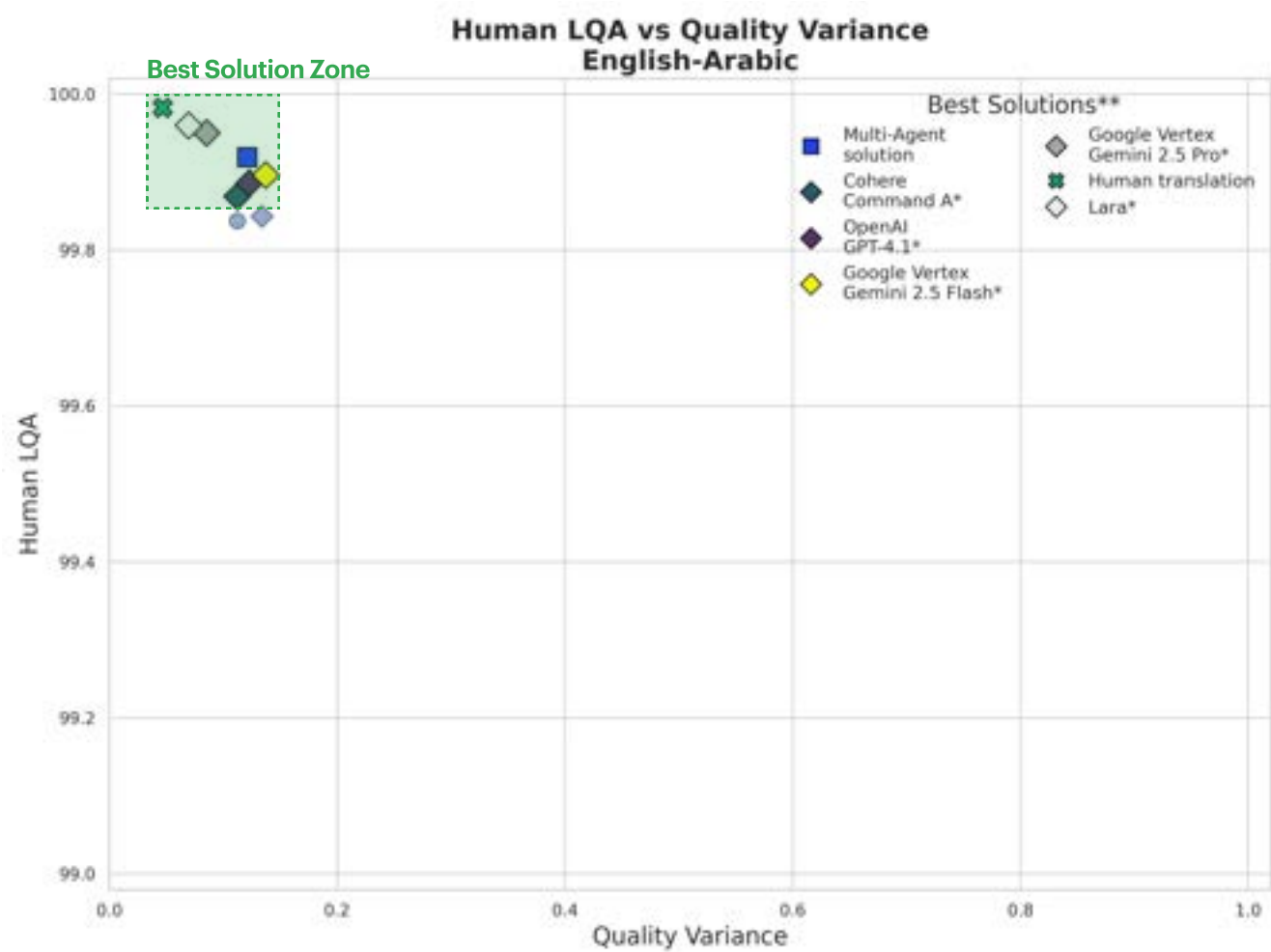
– Baseline – Customized
 – Human Translation – Multi-Agent

All product names, trademarks, and registered trademarks are the property of their respective owners. All company, product, and service names used in this material are for identification purposes only. Use of these names, trademarks, and brands does not imply endorsement.

6.2 Human LQA Adjustment to Report Requirements

- We adjusted human LQA scores by excluding out-of-scope reviewer feedback, specifically stylistic issues and terminology preferences with multiple valid options.
- This prevents translation solutions from being penalized for valid choices that simply differ from a reviewer's personal preference, creating a more objective comparison baseline.
- This methodological refinement ensures fair evaluation by focusing only on objective problems with translations relevant to our requirements.

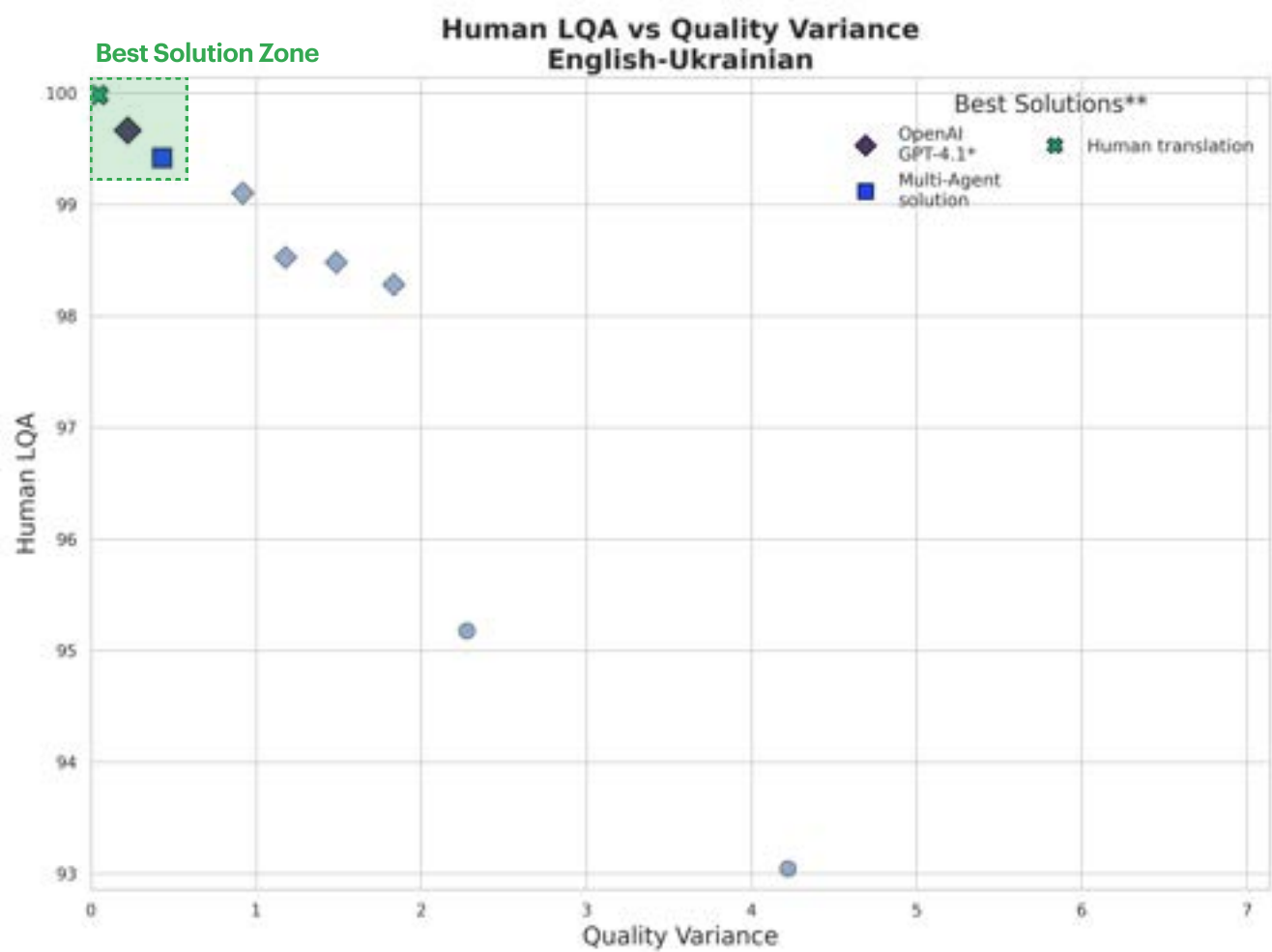
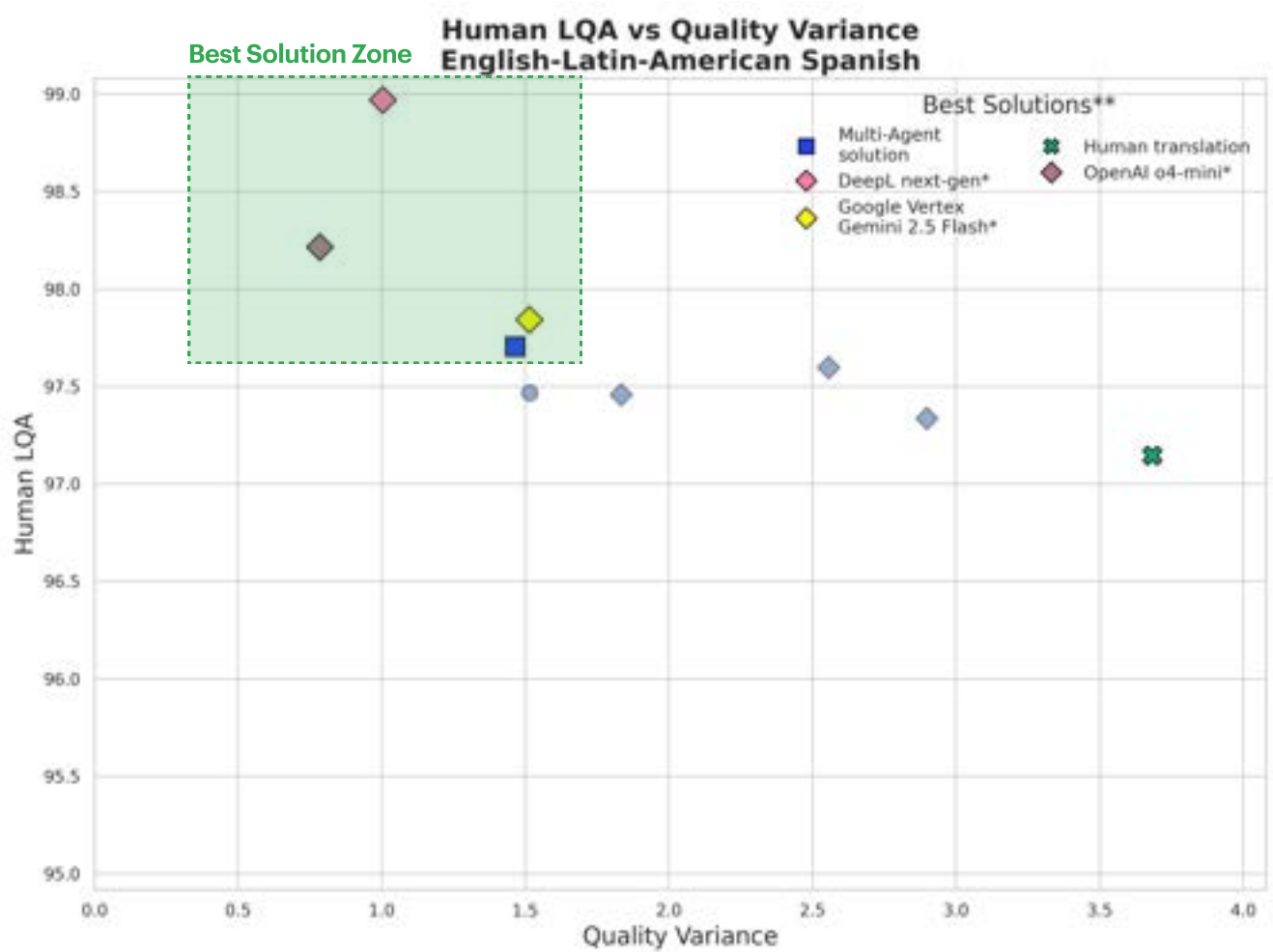
6.3 Human LQA vs Quality Variance Charts (1/2)



- – NMT
- ◆ – LLM
- ✕ – Human translation
- * – Customized

** Human translation is shown on all charts for reference

6.3 Human LQA vs Quality Variance Charts (2/2)



- – NMT
- ◆ – LLM
- ✱ – Human translation
- * – Customized

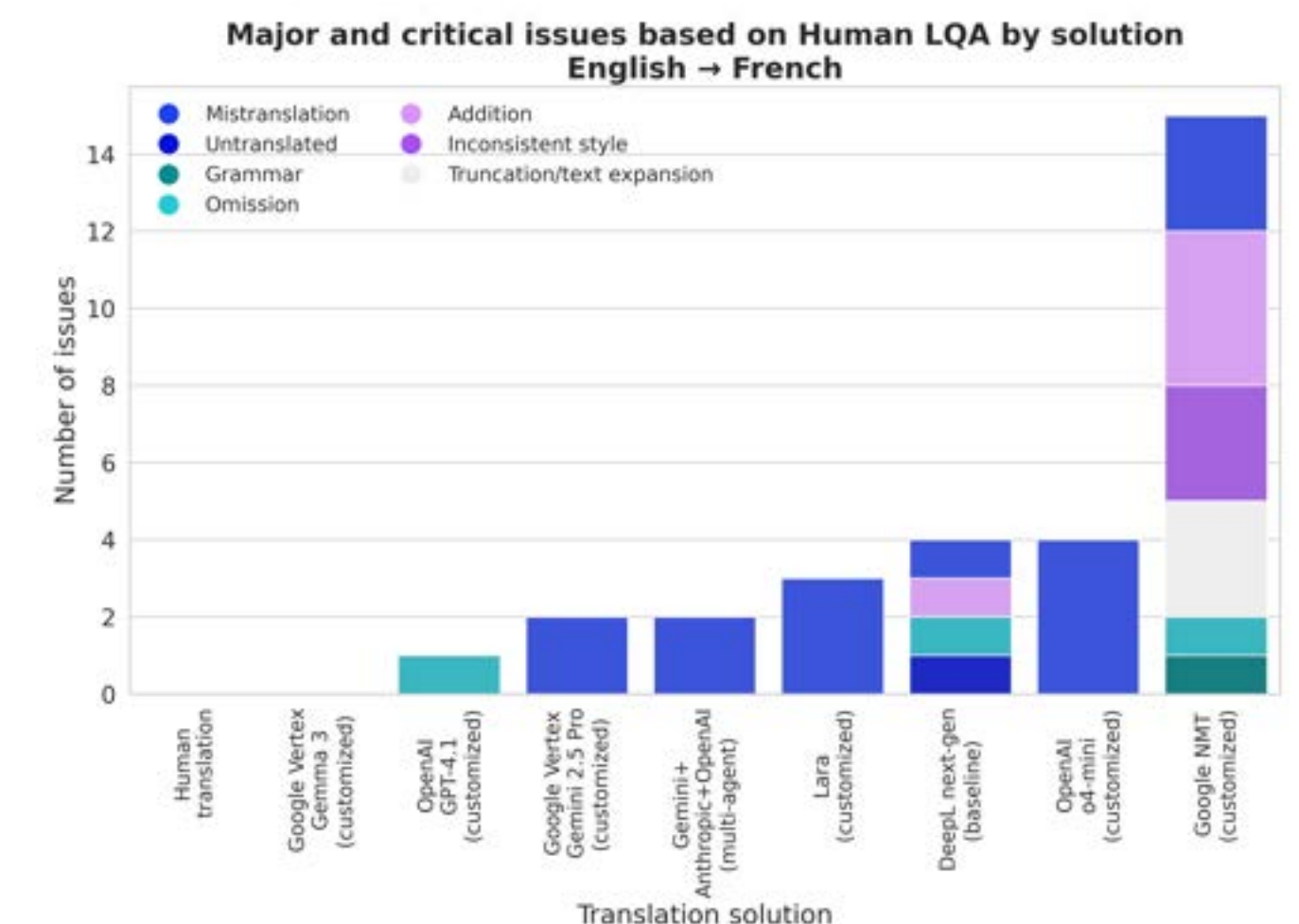
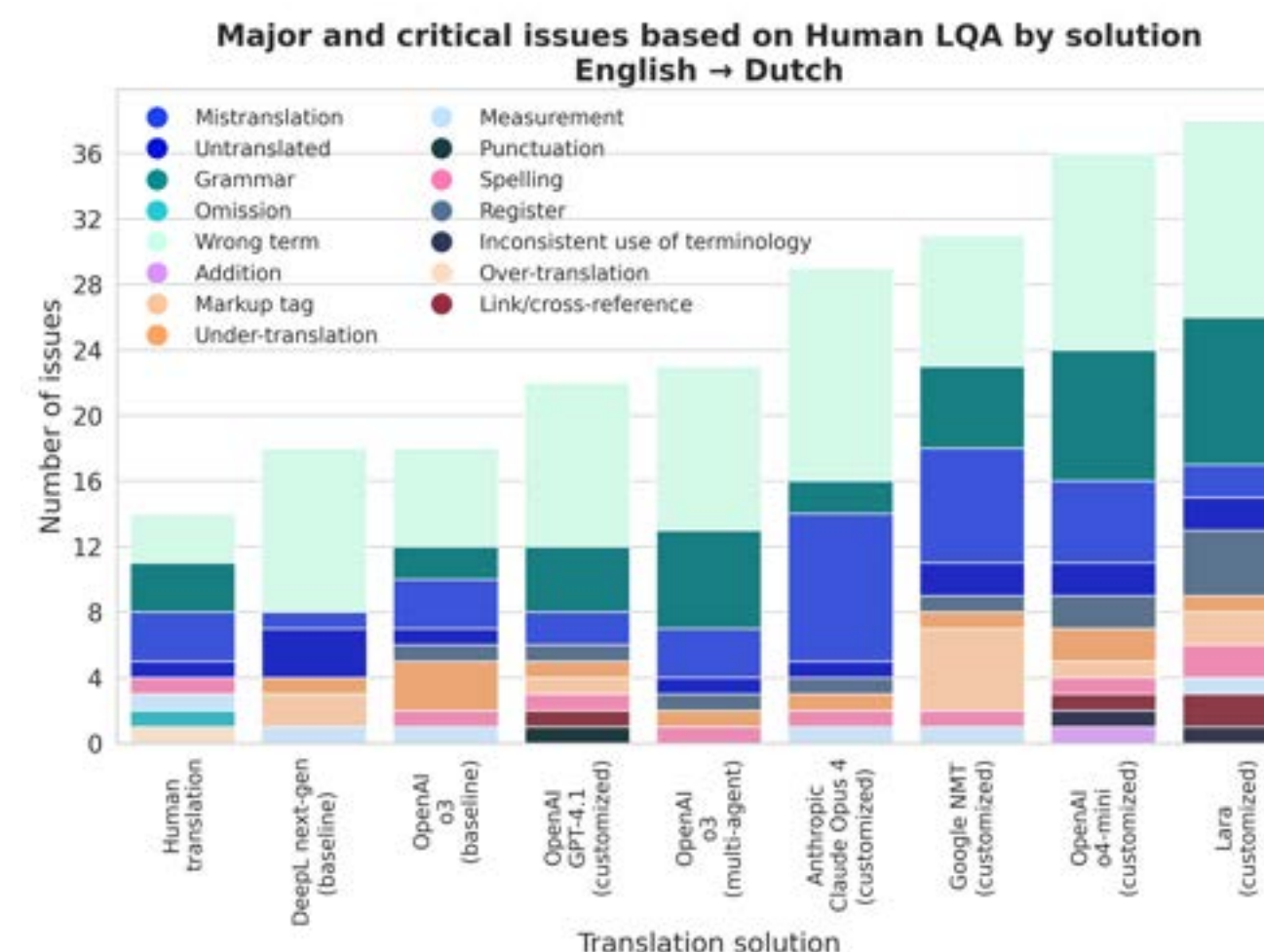
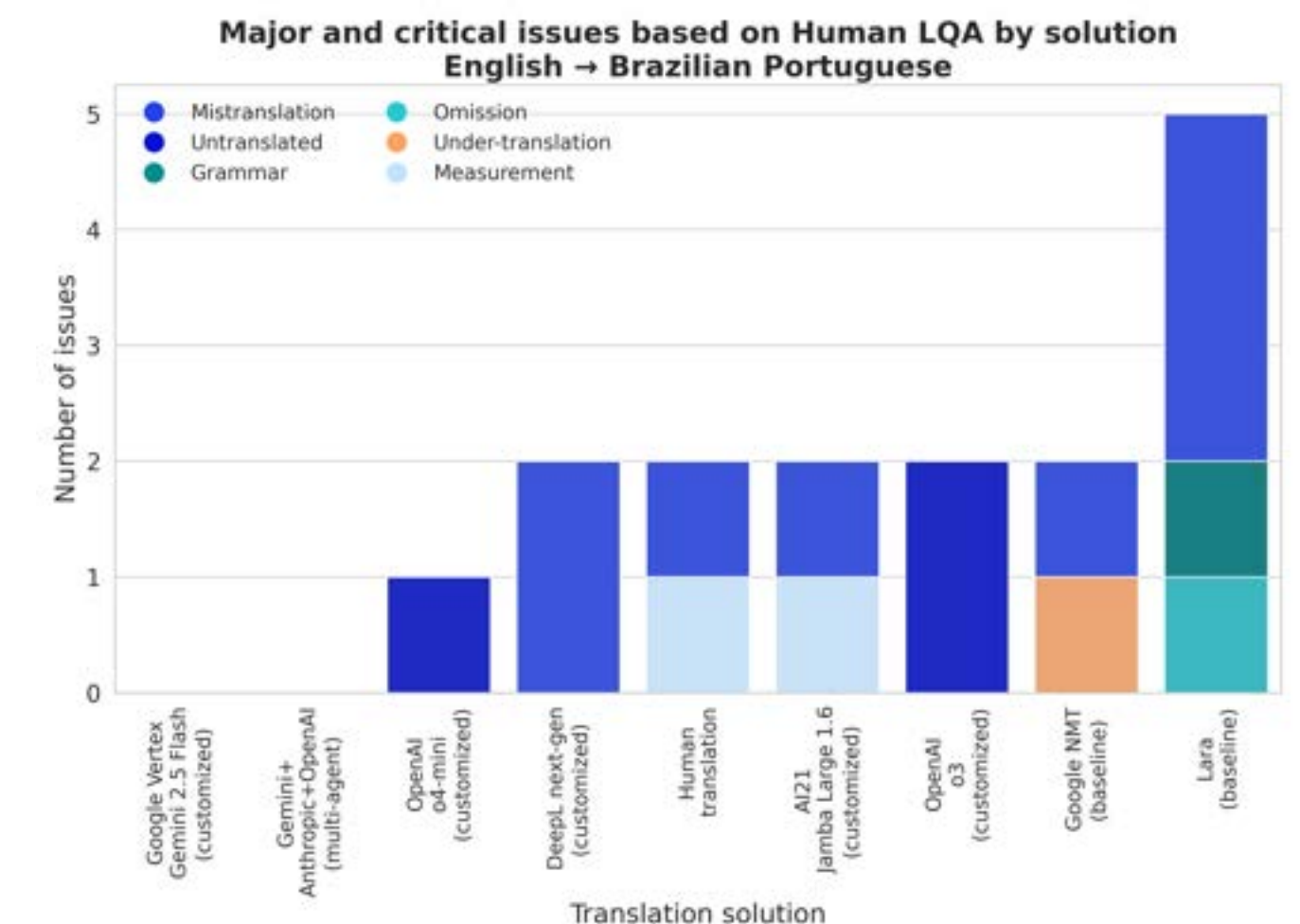
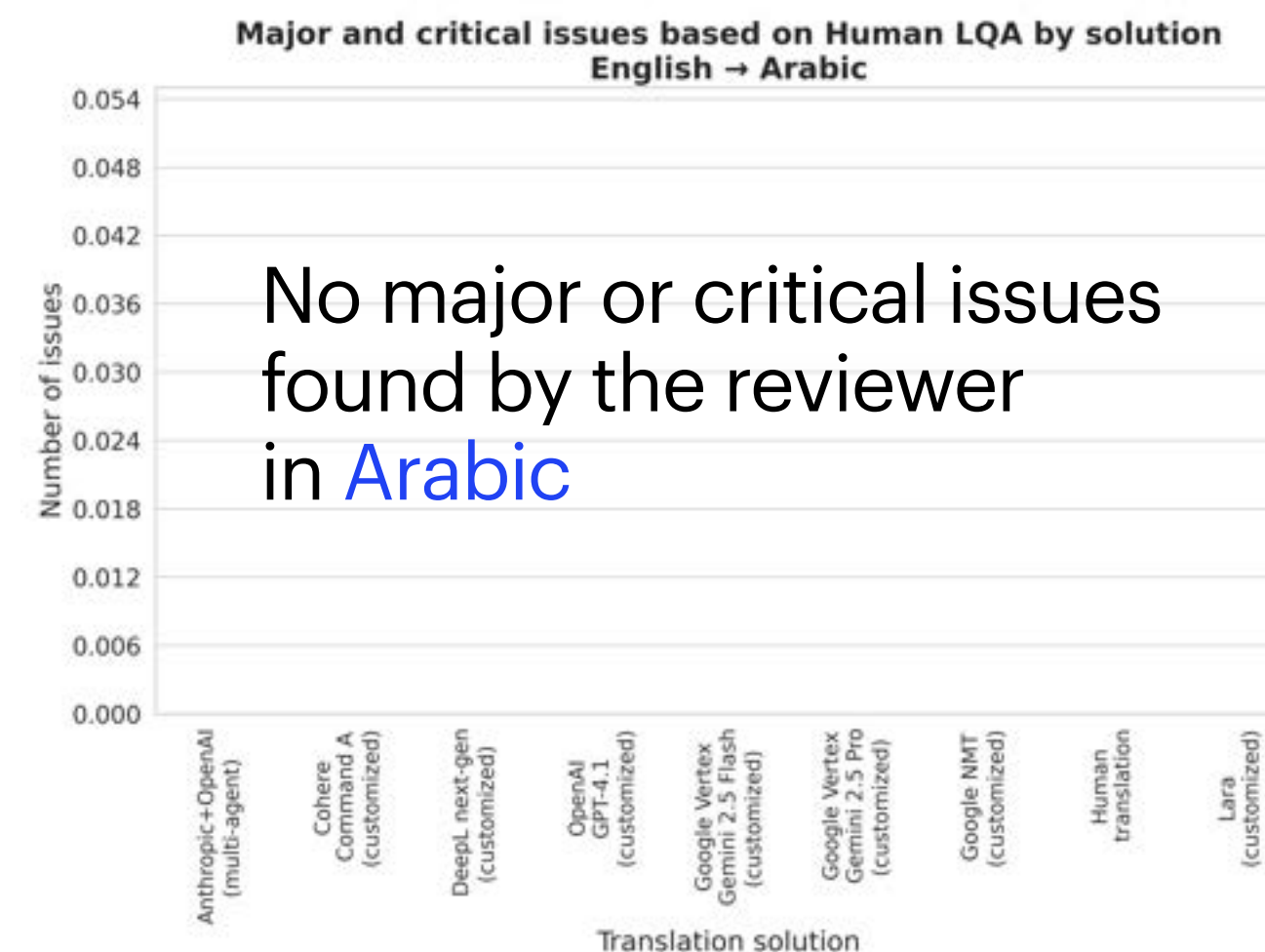
** Human translation is shown on all charts for reference only

6.4 The Human-AI Translation Gap Has Virtually Disappeared

- [Multi-agent solution](#) shows remarkable performance, achieving top-tier results in several language pairs, demonstrating that ensemble approaches can match or exceed single-solution performance.
- The gap between human translation and automated solutions is nearly non-existent in most language pairs.
- [DeepL next-gen](#) performs solidly in [Dutch](#) and [Spanish](#), while [Lara](#) performed among the best in [Arabic](#), [German](#), and [Italian](#), showcasing the benefits of LLMs fine-tuned for translation.
- [Asian languages \(Japanese, Korean, Chinese\)](#) show higher quality variation, indicating highly inconsistent performance across different text types, and fewer solutions with solid performance.
- [European languages](#) generally have more diverse solution representation among the best, with [Anthropic](#), [OpenAI](#) and [Google Vertex](#) models having the highest quality and less variability between texts.
- [Arabic](#) is scored quite sparingly by the reviewer, showcasing high quality among most solutions; there are only minor mistranslation and terminology issues.
- [Ukrainian](#) shows the widest quality spread among different solutions, with NMT engines failing to achieve the near-human translation quality that [Multi-Agent solution](#) and [OpenAI's GPT-4.1](#) come close to; mid-resource languages are notably harder for traditional MT engines.

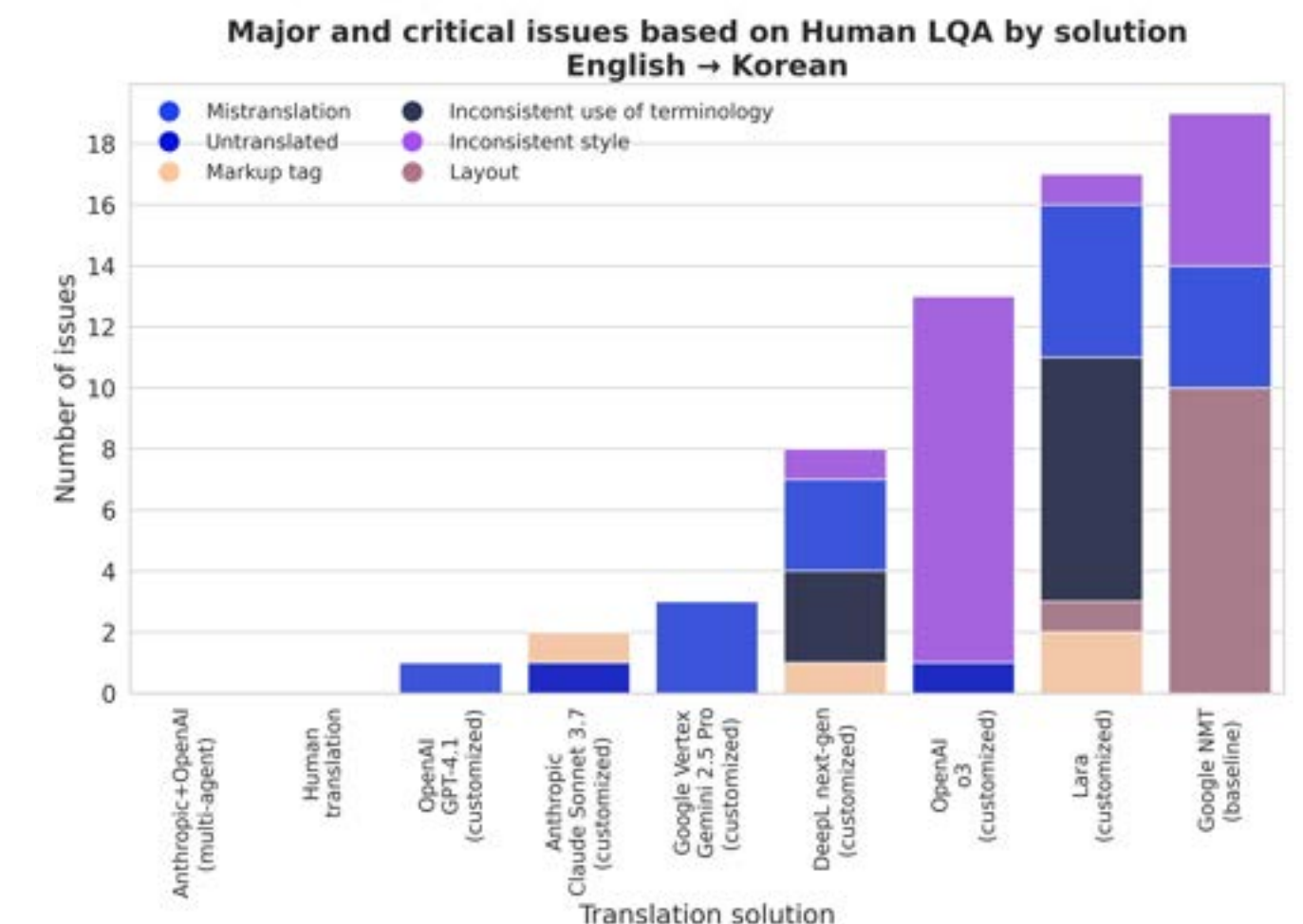
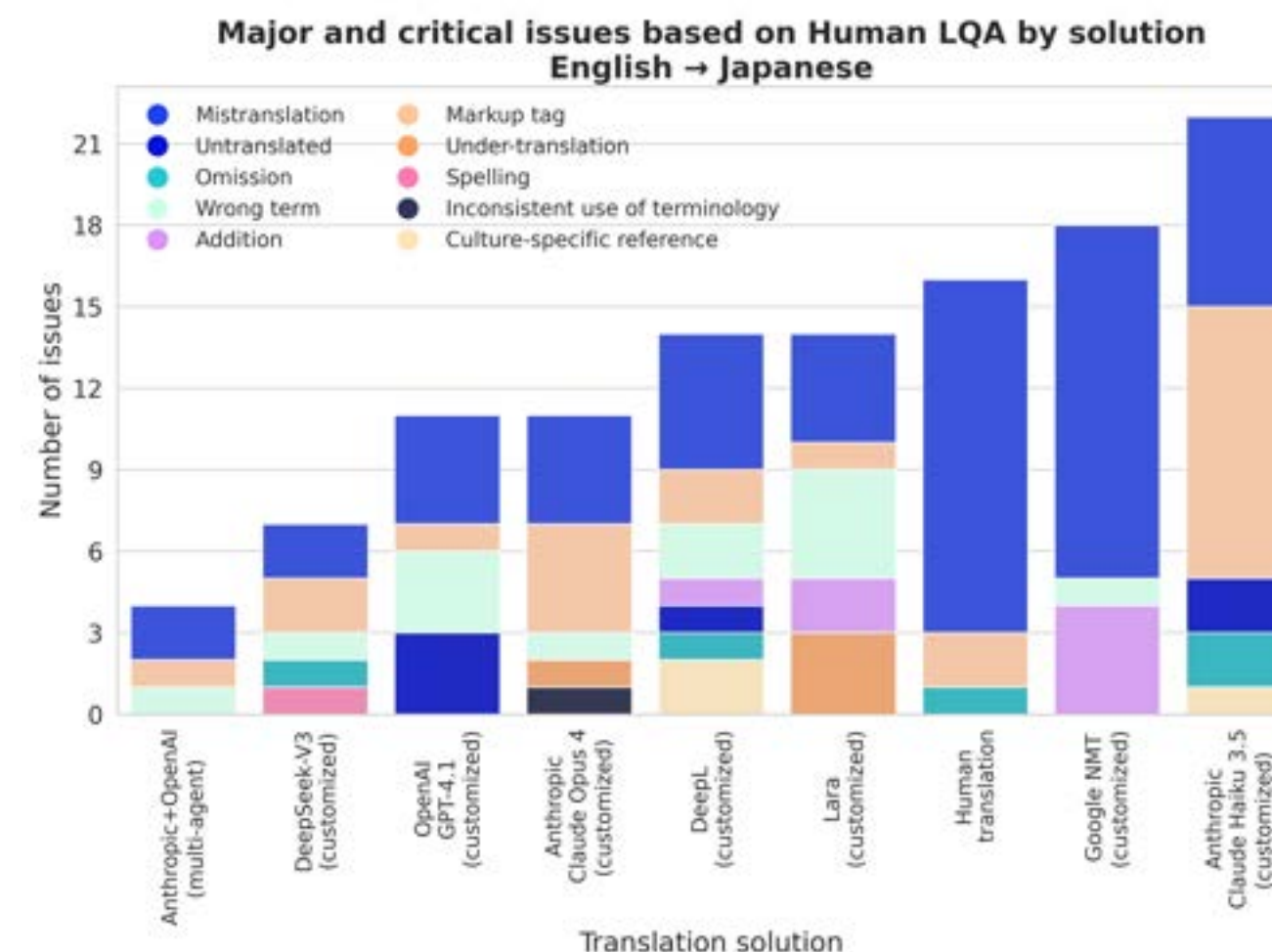
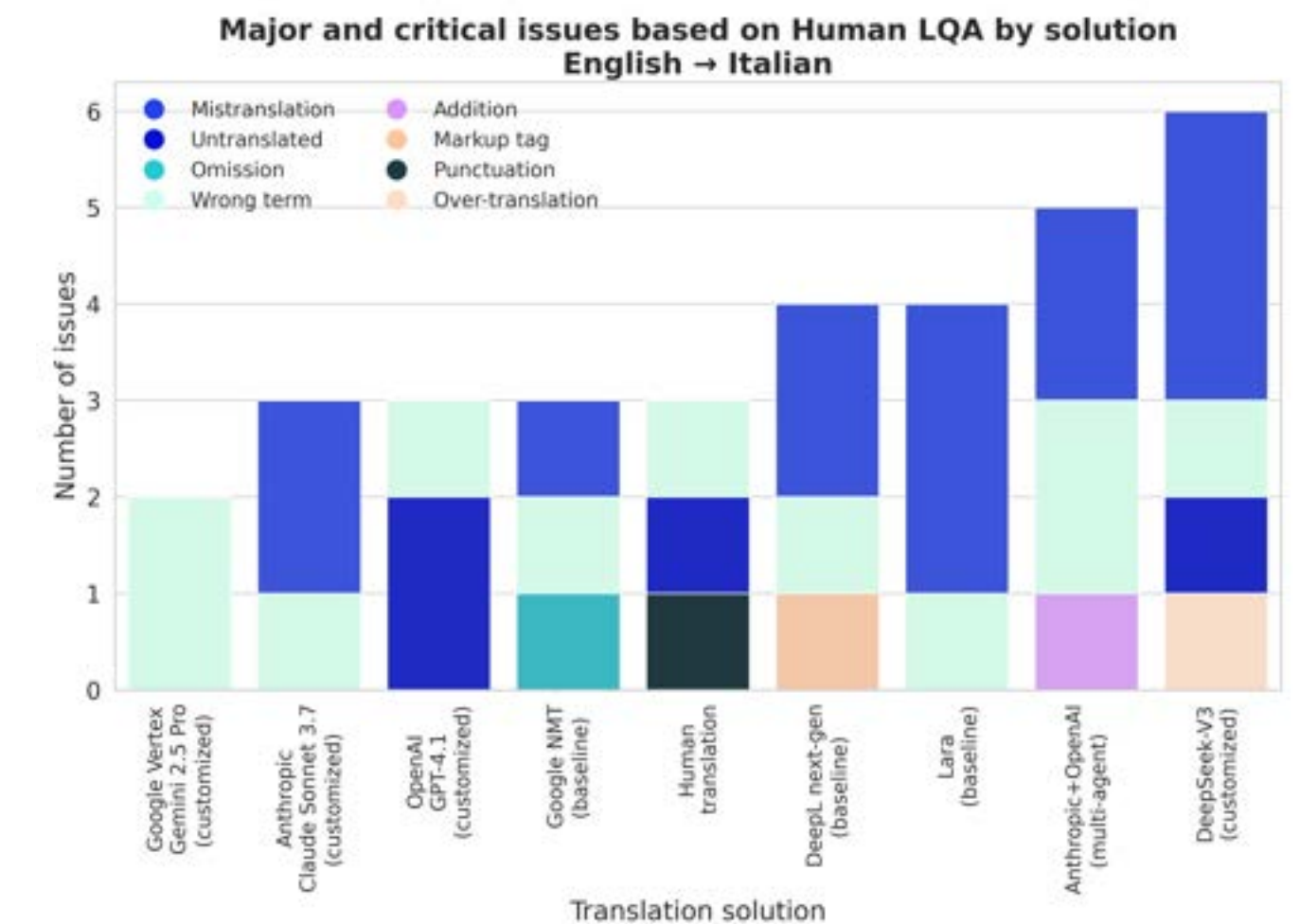
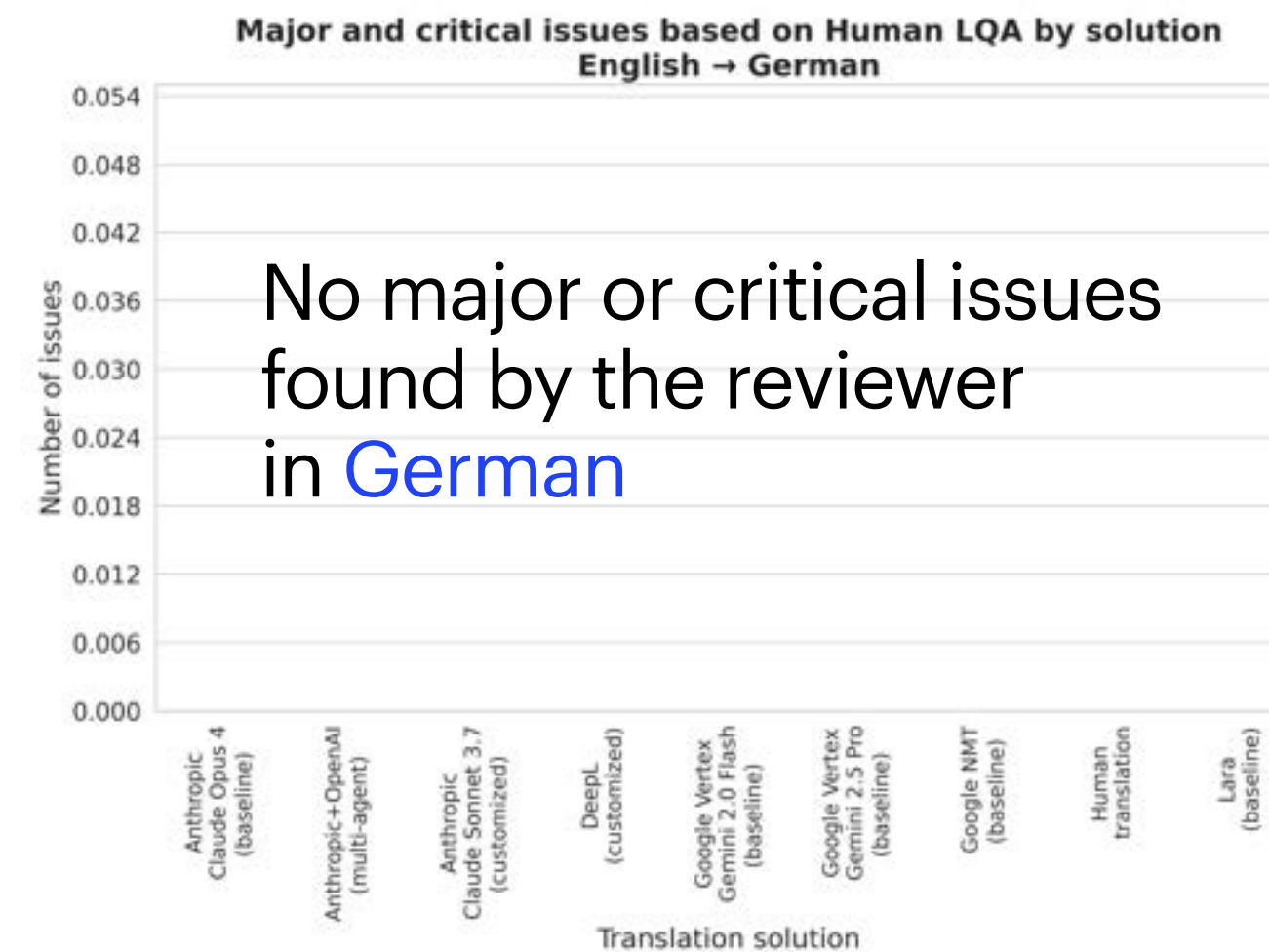
6.5 Error Analysis Reveals Language-Specific Translation Challenges

- The reviewer did not identify any major or critical issues in **Arabic**, having only found 1 to 6 minor grammatical issues per solution.
- In **Brazilian Portuguese**, the most outstanding critical error, seen both in human translation and among several solutions, is an incorrect measurement adaptation from the imperial to the metric system.
- **Dutch** showcases high error counts, with grammar issues and mistranslations dominating. Complex error patterns include markup tags, spelling, and register issues, indicating **Dutch** remains challenging for translation.
- Engine performance varies in **French** with most solutions achieving good results except **Google NMT** showing significantly higher errors, primarily due to untranslated content and truncation issues.



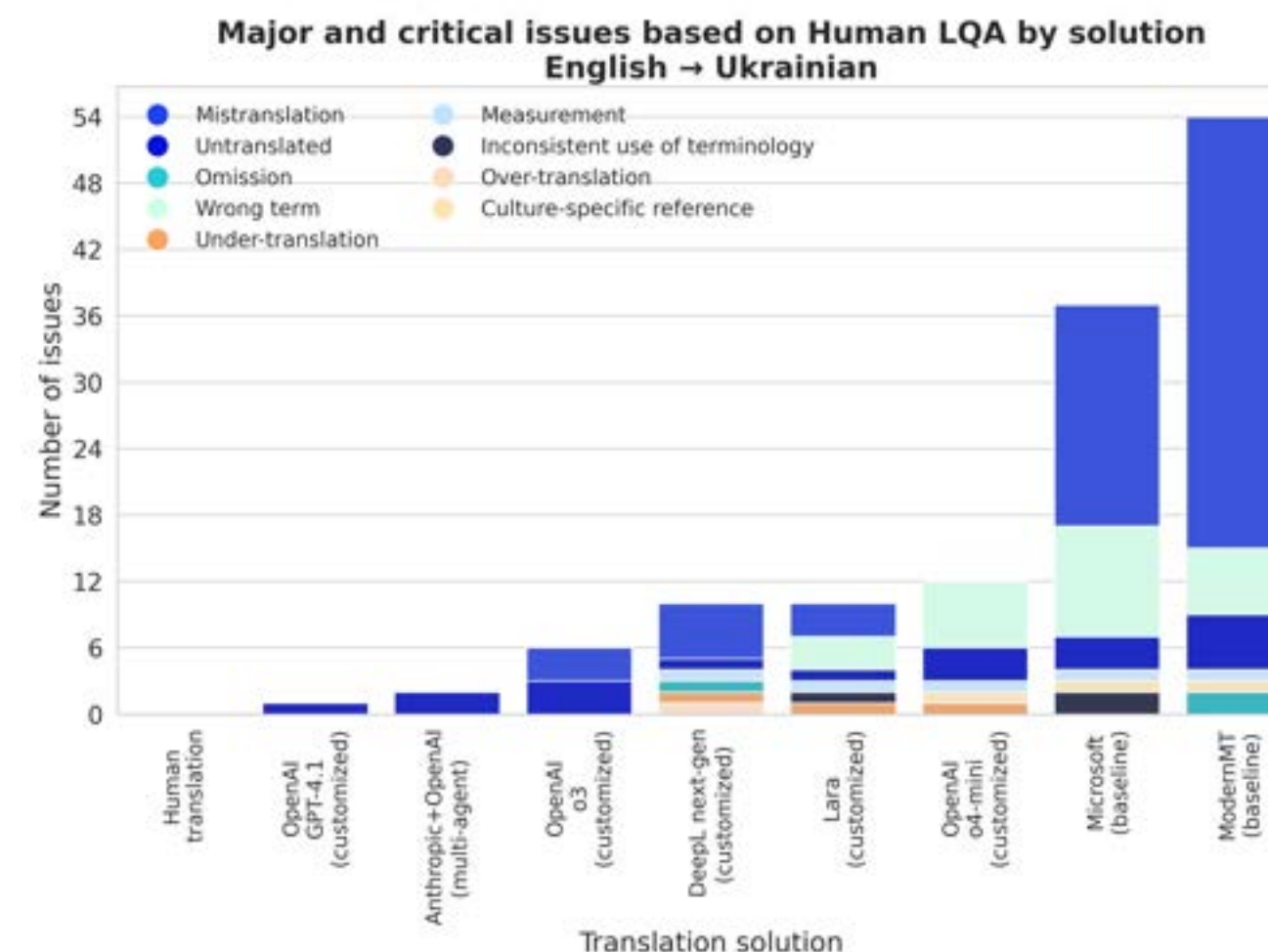
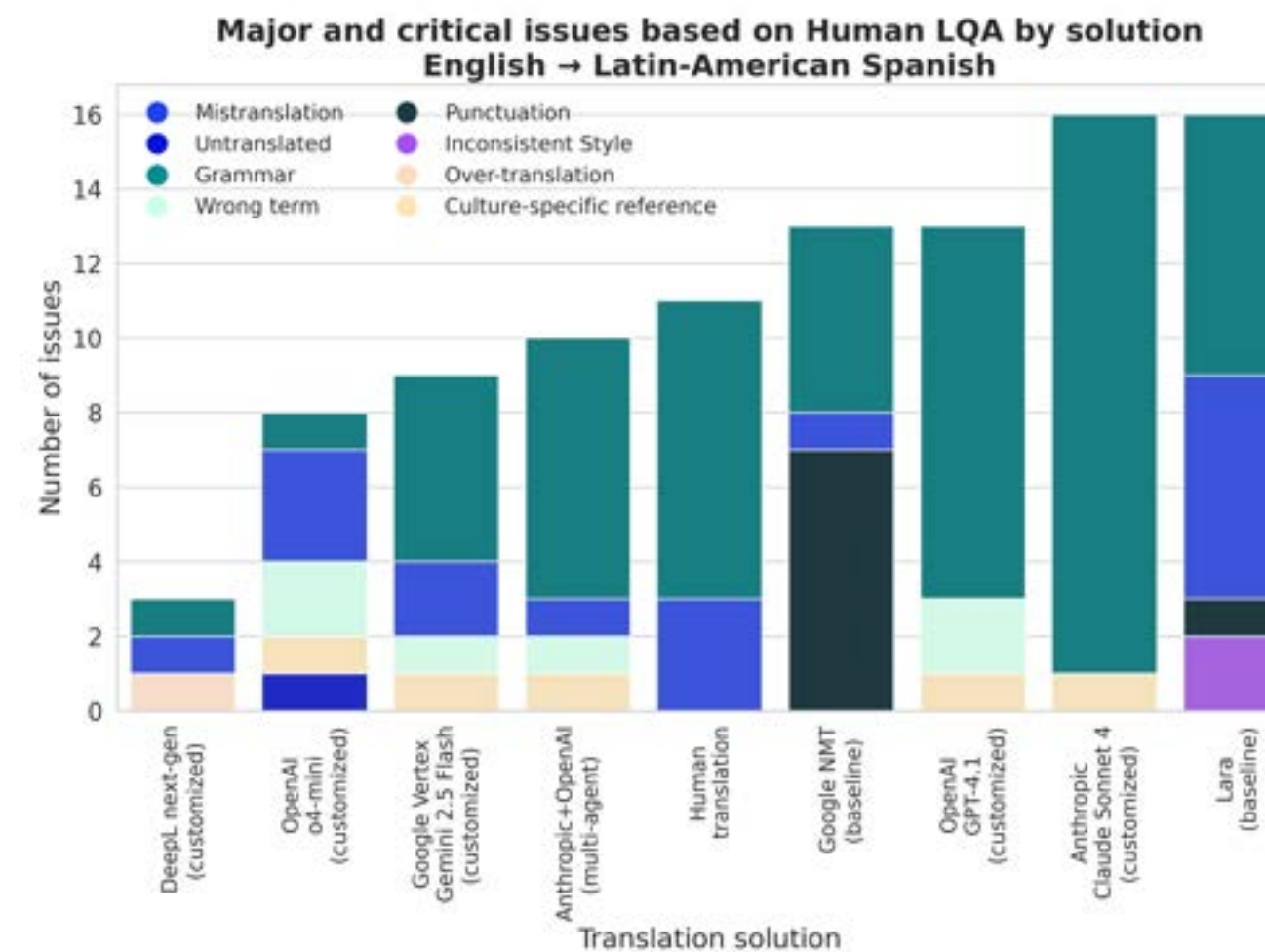
6.6 Human LQA Highlights Language Disparities

- In **German**, no major or critical errors were detected across the board, indicating solid quality from all solutions.
- Solutions show moderate error rates in **Italian**, with mistranslation and wrong term selection being the primary issues, while punctuation and markup problems also appear frequently across multiple engines.
- In **Japanese**, error counts range widely with **Multi-Agent** solutions performing best, while **human translation** shows significantly higher error counts with notable markup errors.
- Solutions struggle with consistently high error rates in **Korean**, where **Multi-Agent** and **human translation** perform with no major or critical errors, while **Lara** and **Google NMT** face significant challenges with terminology and layout.



6.7 Ukrainian Stands as MT's Biggest Challenge While Chinese Excels

- In [Simplified Chinese](#), all solutions achieve exceptional performance with zero major or critical errors detected, demonstrating that despite being a complex language, current engines can successfully process translation requirements with minimal customization.
- Error rates differ in [Latin-American Spanish](#) with [DeepL next-gen](#) performing best (3 errors), while [Claude Sonnet 4](#) and [Lara](#) show the highest error counts, dominated by grammar issues across most solutions.
- In [Ukrainian](#), solutions face extreme challenges with error counts ranging from 1 to 54 issues, where [human translation](#), [GPT-4.1](#) and [Multi-Agent](#) excel with no to minimal errors, while [Microsoft](#) and [ModernMT](#) struggle severely, primarily due to massive mistranslations and untranslated parts of segments.



6.8 Best Solutions per Language Pair Based on Human LQA

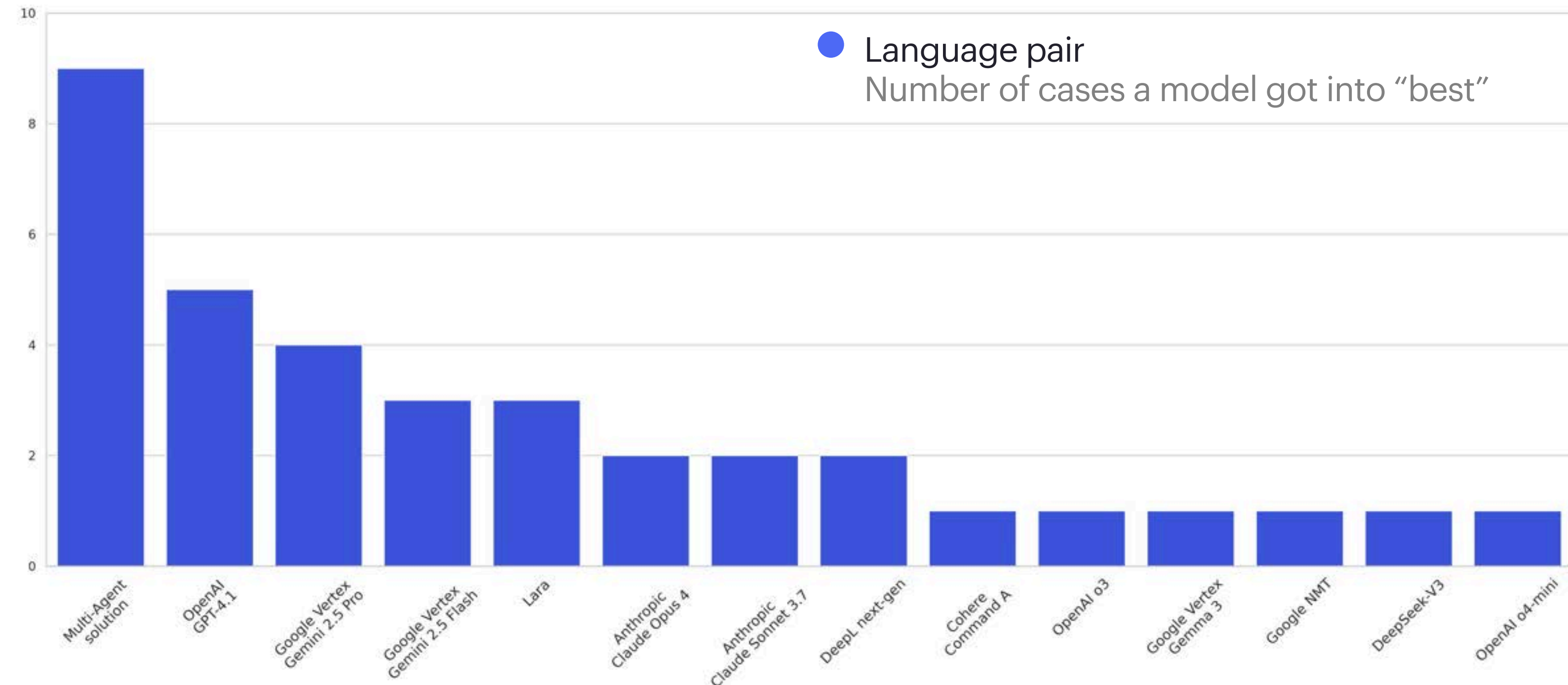
- For each language we average those scores and build a “safety band” around the mean and around the variance (83% confidence).
- A solution is only considered to be “best” if both its quality and variance intervals overlap with the highest scoring solution (see Section 3.7).
- Real-time best solutions based on Human LQA can be found in [Appendix C](#).

⚙️ – Customized 🤖 – Multi-Agent 📄 – Baseline

Best solutions by Human LQA						
	AI Anthropic	Cohere	DeepL	DeepSeek	Google	Multi-Agent
			OpenAI	Translated		
en-ar	Cohere Command A	Gemini 2.5 Flash	Gemini 2.5 Pro	GPT-4.1	Multi-Agent Solution	Translated Lara
en-de	Claude Opus 4	Claude Sonnet 3.7	Gemini 2.5 Pro	Translated Lara		
en-es	DeepL next-gen	Gemini 2.5 Flash	Multi-Agent Solution	OpenAI o4-mini		
en-fr	Google Gemma 3	GPT-4.1				
en-it	Claude Sonnet 3.7	Gemini 2.5 Pro	Google NMT	GPT-4.1	Multi-Agent Solution	Translated Lara
en-ja	DeepSeek-V3	Multi-Agent Solution				
en-ko	Gemini 2.5 Pro	Multi-Agent Solution				
en-nl	Claude Opus 4	DeepL next-gen	GPT-4.1	Multi-Agent Solution	OpenAI o3	
en-pt	Gemini 2.5 Pro	Multi-Agent Solution				
en-uk	GPT-4.1	Multi-Agent Solution				
en-zh	Multi-Agent Solution					

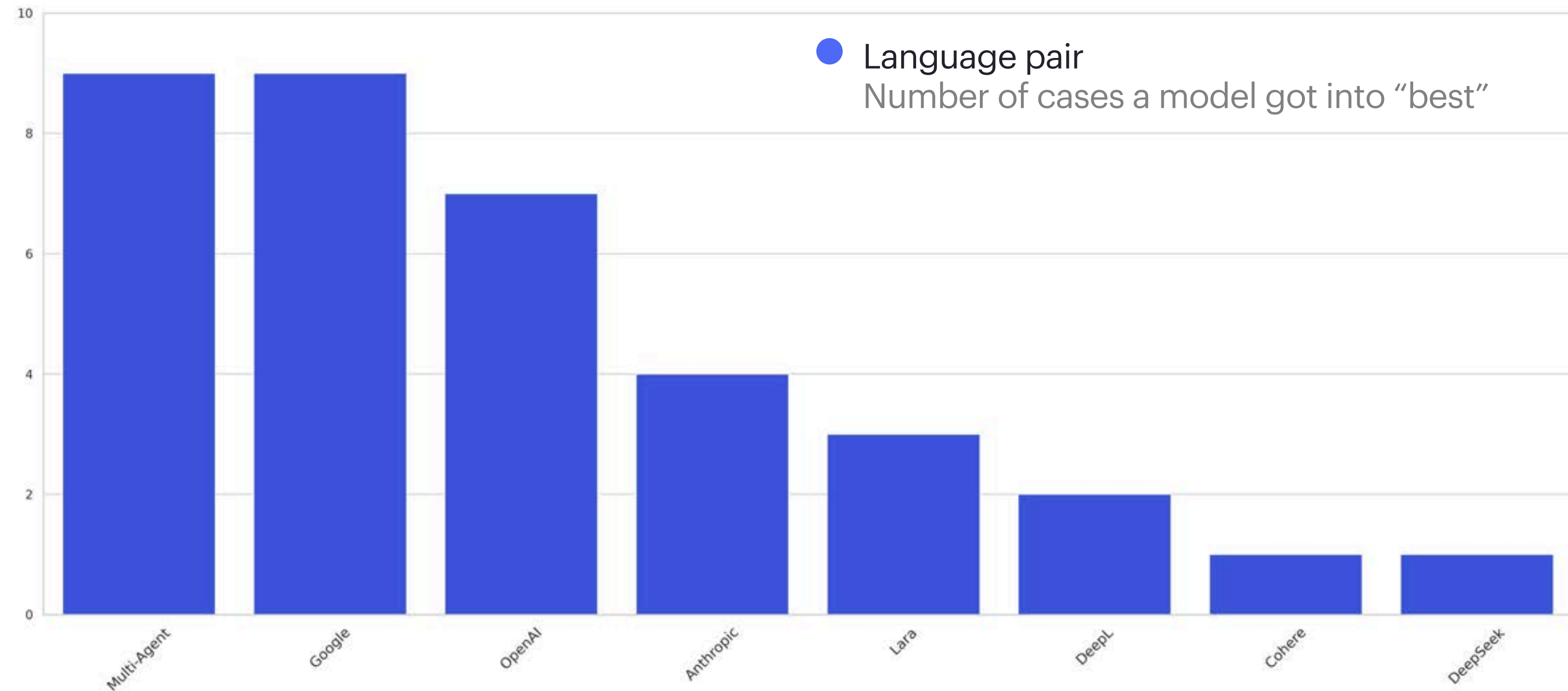
6.9 Multi-Agent Solution Outperforms All Others by a Landslide

- **Multi-Agent solution** achieves 9 "best" ratings out of 11 possible, demonstrating that ensemble approaches significantly outperform individual models when consistency across languages is critical.
- Traditional MT providers and newer LLM solutions cluster at the bottom with only one or two wins each. This suggests that neither legacy expertise nor the latest architectural advances alone can guarantee translation excellence—the future lies in the intelligent combination of models.



6.10 Several Google Solutions Perform On Par With the Multi-Agent Approach

- [Multi-Agent solution](#) and [Google](#) tie at 9 "best" performances each, showing that Google's diverse model portfolio ([Gemma](#), [Gemini](#) family, [Translation LLM](#)) can collectively match the consistency of a single ensemble approach.
- LLMs fine-tuned for translation, [Lara](#) and [DeepL next-gen](#) showcase the new kind of translation solutions, achieving the "best" status in several language pairs.



7. Miscellaneous

7.1 195,928 Language Pairs Across All MT Engines

7.2 LLMs are 10-100 Times Slower than Specialized MT Models

7.3 Changes in Providers' Features

7.4 Translation-Focused LLMs Adopt MT Pricing

7.5 Public Pricing for Model Customization

7.6 Independent Vendors Grow in Numbers with LLMs Leading Innovation

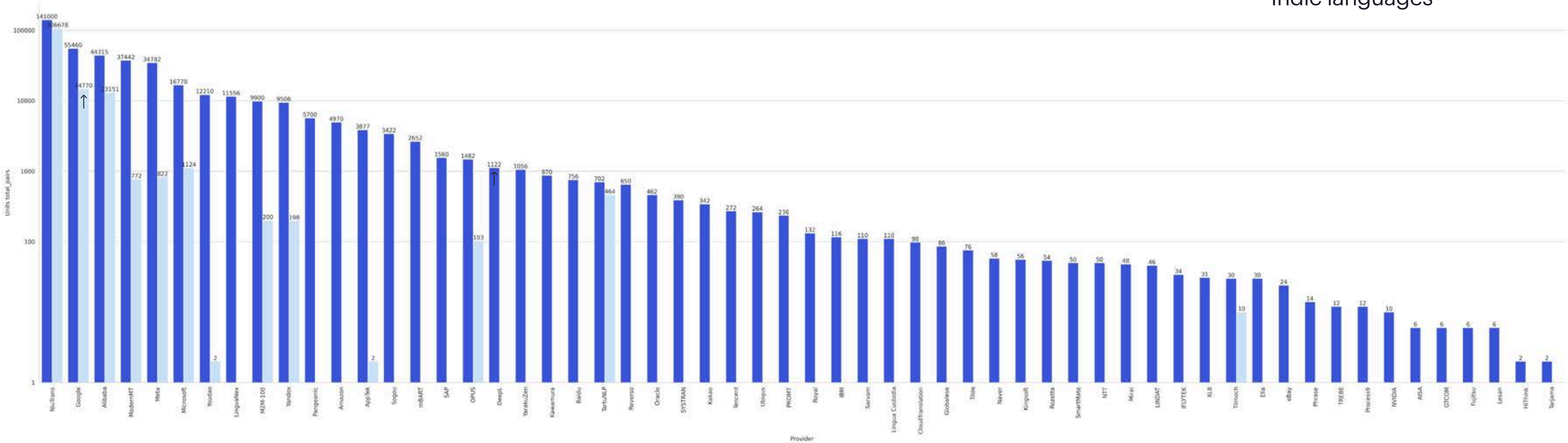
7.1 195,928 Language Pairs Across All MT Engines*

- total language pairs
- unique language pairs
- ↑ language pair growth

From 191,010 in May'24 to 195,928 in June'25

Google has 110 new languages in August'24, 26 of them were new to Intento

Added 2 new providers — [Tilmoch](#), specializing in Turkic languages, and [Sarvam](#), a translation LLM for Indic languages



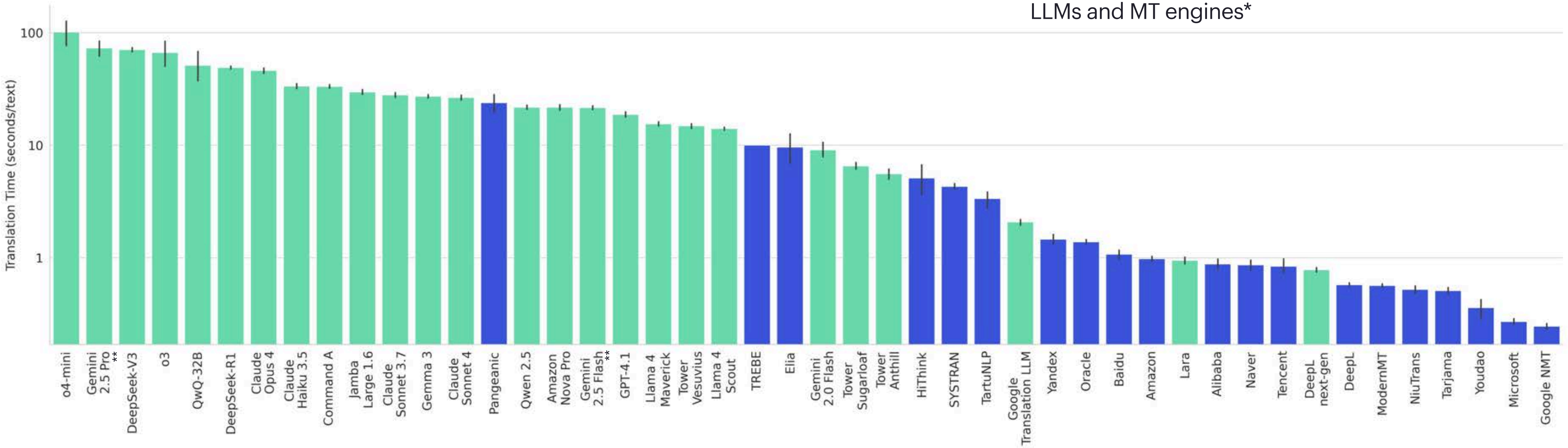
* Where possible, we have checked via API if all language pairs advertised by the documentation are supported and removed the pairs we were unable to locate in the API.
** As advertised (not validated via API).

*** Due to LLMs multilinguality and the nature of the data they were trained on, there is no definitive list of languages or language combinations they support. For now, we do not include LLMs in this slide.

7.2 LLMs are 10-100 Times Slower than Specialized MT Models

- NMT engines
- Large Language Models

Even with a general trend towards latency improvement, there is still a 10–100 times difference in translation time between most LLMs and MT engines*



* Several MT Engines’ speed was affected by limited quotas (Elia, TartuNLP, Pangeanic)

** Gemini 2.5 Pro and Gemini 2.5 Flash have only been tested for latency as Preview versions; the GA models have better latency based on our research testing

7.3 Changes in Providers' Features

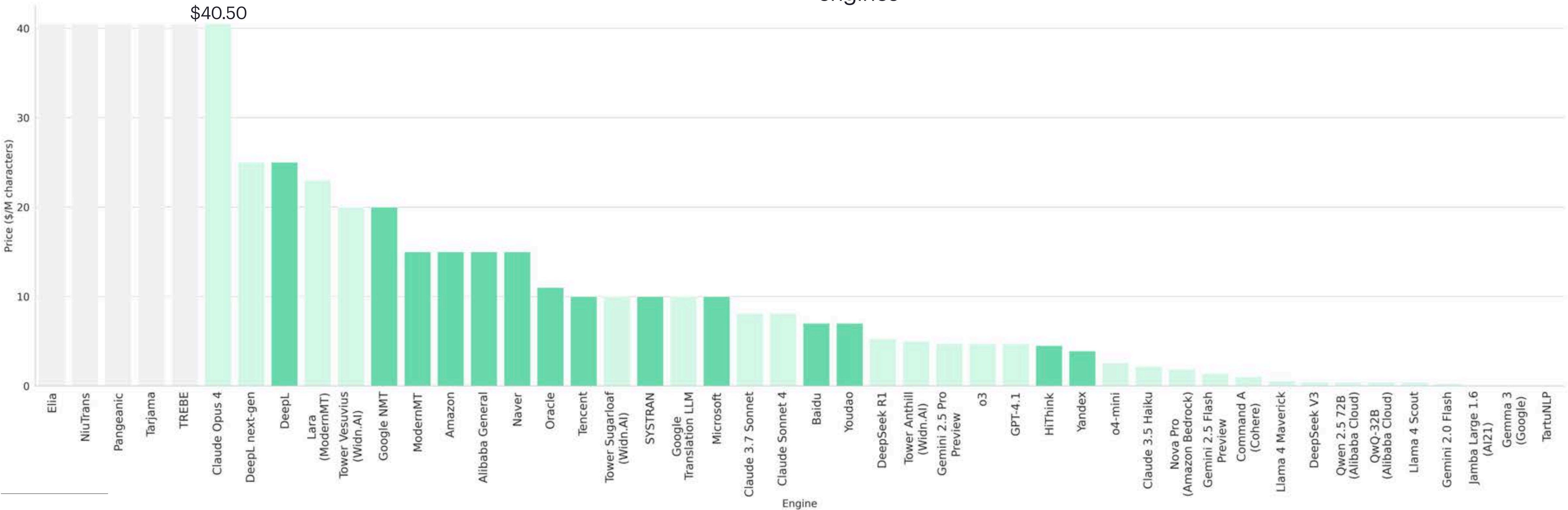
- DeepL [launched](#) a next-generation LLM specifically designed for translation that provides improved quality especially for longer texts, supports all languages from their classic model plus Hebrew and Vietnamese, and maintains the same high enterprise security standards.
- Google Translate [rolled](#) out 110 new languages, adding translation support for more than 614 million speakers (around 8% of the world's population), with about a quarter of the new languages coming from Africa and including major additions like Cantonese, Punjabi (Shahmukhi), and various Indigenous languages.
- Translated [released](#) Lara, an AI-powered translation platform, which delivers more accurate translations than the median professional translator, with plans to achieve "language singularity" by 2025 when it expects to match the quality of the top 1% of professional translators.
- Chinese AI startup DeepSeek [introduced](#) R1 and V3, open-source language models that can nearly match the capabilities of OpenAI's, Meta's and Google's Large Language Models at a fraction of the cost, spending just \$5.6 million powering its base AI model.
- Microsoft [released](#) Microsoft Translator Pro for enterprise users in January 2025, featuring customized phrasebooks, expanded language coverage, and availability in US Government cloud, marking a shift toward premium enterprise services.

7.4 Translation-Focused LLMs Adopt MT Pricing

- stock provider price (\$/1 MLN characters)
- LLM price (\$/1 MLN characters)
- price by request

Most Large Language Models are priced 10-100 times lower than traditional Machine Translation engines

Newest LLMs fine-tuned for the Machine Translation task are priced similarly to MT engines



* Open Source Engines

** Prices for LLMs are converted with an estimation of 2.83 characters per token on average.

*** Prices provided herein are based on the publicly listed prices at the time of the analysis. Actual prices may vary depending on a variety of factors, including your geographical location and any customary discounts. It is always recommended to contact the vendor directly for the most accurate and up-to-date pricing information.

7.5 Public Pricing for Model Customization

	Amazon	Gemini 2.0 Flash	Google v3	GPT-4.1	LLaMA models	Microsoft	ModernMT Human in the Loop	SYSTRAN
Customization (\$)	free	by request	\$45/hour, \$300 max	\$8.83/M characters*	free	\$10/M source + target characters (max. \$300)	free	by request
Hosting/month (\$)	200 GB free parallel data storage. \$0.023/GB per month for excess data	by request	free	free	free	\$10	free	by request
Translation (\$/M characters)	\$60	by request	\$80	\$7.04*	free	\$40	\$50	by request

* Prices provided herein are based on the publicly listed prices at the time of the analysis. Actual prices may vary depending on a variety of factors, including your geographical location and any customary discounts. It is always recommended to contact the vendor directly for the most accurate and up-to-date pricing information.

** Prices are converted with an estimation of 2.83 characters per token.

7.6 Independent Vendors Grow in Numbers with LLMs Leading Innovation

Commercial

AISA, Alibaba, Amazon, Apptek, Baidu, CloudTranslation, DeepL, Elia, Fujitsu, Globalese, Google, GTCOM, IBM, iFlyTek, RoyalFlush, Lesan, Lindat, Lingua Custodia, Lingvanex, Kantan, Kawamura / NICT, Kingsoft, Masakhane, Microsoft, Mirai, ModernMT, Naver, Niutrans, NTT COTOHA, Omniscien, Pangeanic, Prompsit, PROMT, Process9, Reverso, Rozetta, RWS, [Sarvam](#), SAP, Sogou, Systran, Tarjama, Tencent, Tilde, [Tilmoch](#), Ubiquis, TREBE, XL8, Yandex, YarakuZen, Youdao

Preview / Limited

eBay, Kakao, QCRI, Tarjama, Birch.AI

Open Source Pretrained

TartuNLP, NeMo by NVIDIA, NLLB by Meta AI, M2M-100, mBART, OPUS

Large Language Models

01.AI, AI21, [Aleph Alpha](#), Alibaba, AllenAI, [Amazon AWS](#), AMD, Anthropic, BAAI, Baichuan, [Baidu](#), BAIR, [Barcelona Supercomputing Center](#), BigScience, Cerebras, Cohere, Databricks, Deci, [DeepSeek](#), Eleuther AI, [EU Commission](#), Google, HuggingFace, [IBM](#), LianjiaTech, LLM360, LLMZoo, Meta, Microsoft, Mistral, [NVIDIA](#), OpenAI, Preferred Networks, Salesforce, [Snowflake Arctic](#), StabilityAI, StatNLP, [Tencent](#), TII, [Unbabel](#), xAI

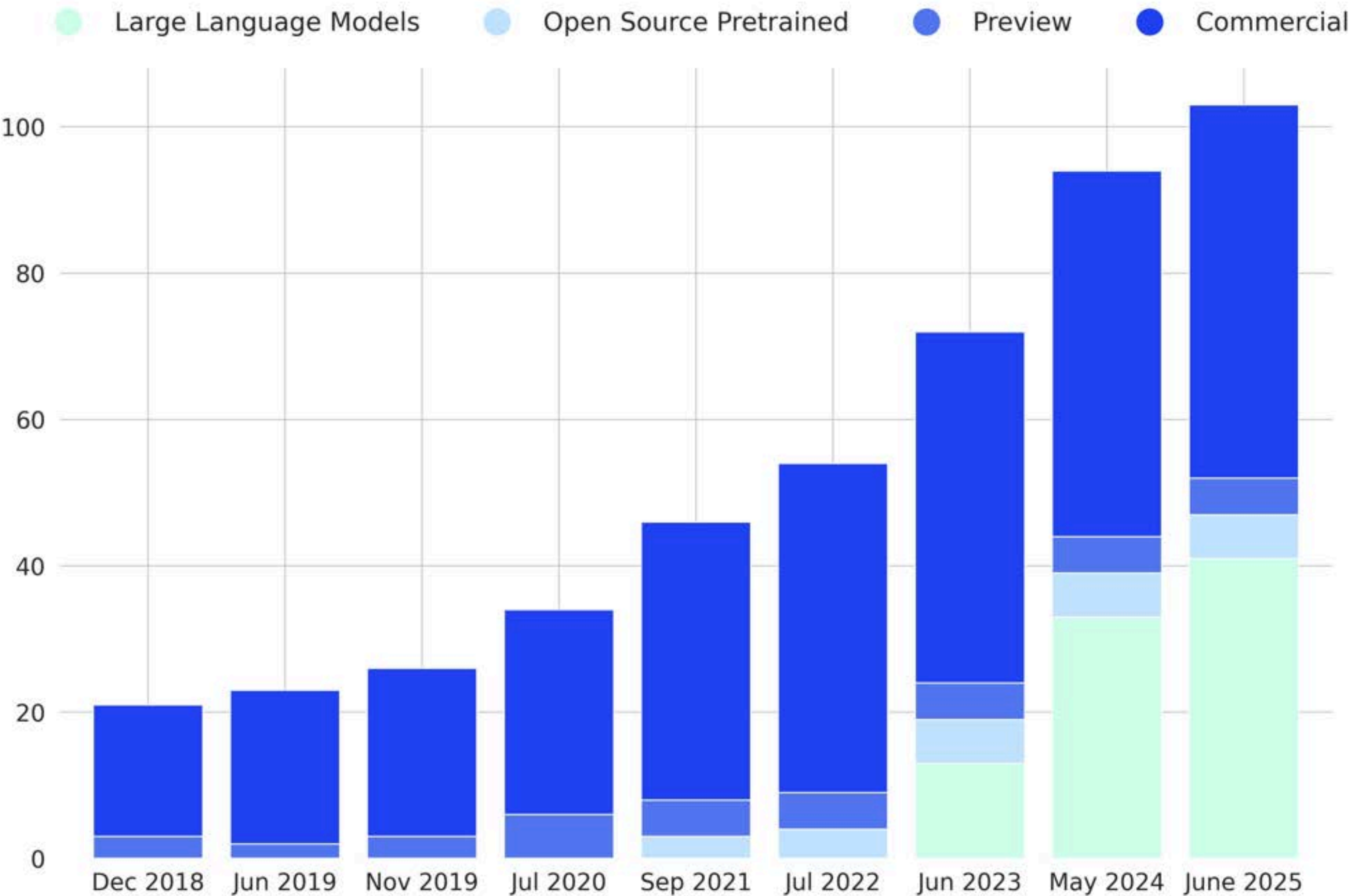
The new engines are highlighted in blue

51

5

6

41



8. Takeaways

8.1 Key Conclusions

8.2 Intento – Machine Translation and multilingual AI agents platform for global businesses

8.3 Unlock consistent and authentic translation across all touchpoints with an all-in-one, scalable platform

8.4 Intento's requirements-driven translation

8.5 Meet all your translation requirements with AI agents

8.6 Intento LQA: actionable, tailored linguistic quality assessment for AI-first localization

8.1 Key Conclusions (1/2)

The raw MT quality ceiling is rising rapidly

With top solutions achieving 99+ Intento LQA scores and minimal human-evaluated errors, we're approaching a plateau where improvements will focus on subtle nuances rather than fundamental accuracy.

Requirements-based customization is no longer optional

Analysis of terminology, tone, formatting, and consistency requirements shows baseline systems produce 10-15 errors per text, while customized solutions only have 0-2 errors. This 5-10x improvement in accuracy makes customization essential for any professional translation workflow.

Ukrainian remains machine translation's greatest challenge within this report

With quality variance ranging from 1 to 7 points, Ukrainian stands alone as the most difficult language for current systems. This persistent challenge, coupled with the language's mid-resource status and complex grammar, suggests that specialized approaches or significant model improvements are needed for languages with similar profiles.

Multi-agent workflows achieve near-human translation quality

Multi-agent systems combining several agents for glossary integration, tone adjustment, and post-editing consistently rank in the best solutions category across [Brazilian Portuguese](#), [Dutch](#), [Italian](#), [Japanese](#), [Korean](#), [Latin-American Spanish](#) and [Simplified Chinese](#), offering a reliable path to high-quality automated translation today.

Customization transforms mid-tier engines into top performers

Engines with baseline Intento LQA scores between 96-98 show the most dramatic improvements from customization, often gaining 2-3 points to reach 98-99 range. Models like [Cohere Command A](#), [Google Vertex Gemini 2.5 Flash](#), and [DeepL](#) benefit substantially from terminology and tone customization.

Terminology accuracy demands multi-agent workflows or customization

For critical terminology (medical, legal, technical), only human translation, multi-agent workflows, and customized LLMs would deliver acceptable results.

8.1 Key Conclusions (2/2)

LLMs are 10-100x slower than NMT but they are getting there

While models like [Google Vertex Gemini 2.0 Flash](#), [Widn.AI Tower](#) models, and translation-focused LLMs ([Lara](#) from [Translated](#), [DeepL next-gen](#)) show notable latency improvements compared to LLM performance in 2024, the fundamental gap persists as LLMs process at 10-100 seconds per text versus NMT's sub-second speeds.

The MT landscape reaches unprecedented scale with 195,928 language pairs

With providers like [Google](#) adding 110 new languages and specialized vendors like [Tilmoch](#) and [Sarvam](#) targeting underserved markets, machine translation is rapidly closing the gap for low-resource languages.

Translation requirements expose clear capability tiers

When evaluated against specific business requirements (terminology, tone, consistency, formatting), distinct tiers emerge: customized LLMs and multi-agent workflows achieving near-perfection, customized NMT engines with solid performance, and baseline systems typically producing the most errors. This 10x error rate difference directly impacts post-editing costs.

Translation-focused LLMs adopt competitive pricing strategies

[Lara](#), [DeepL next-gen](#), and [Tower](#) models price competitively with traditional NMT engines, suggesting the industry is moving toward purpose-built LLMs delivering superior quality without the premium pricing historically associated with NMT engines.

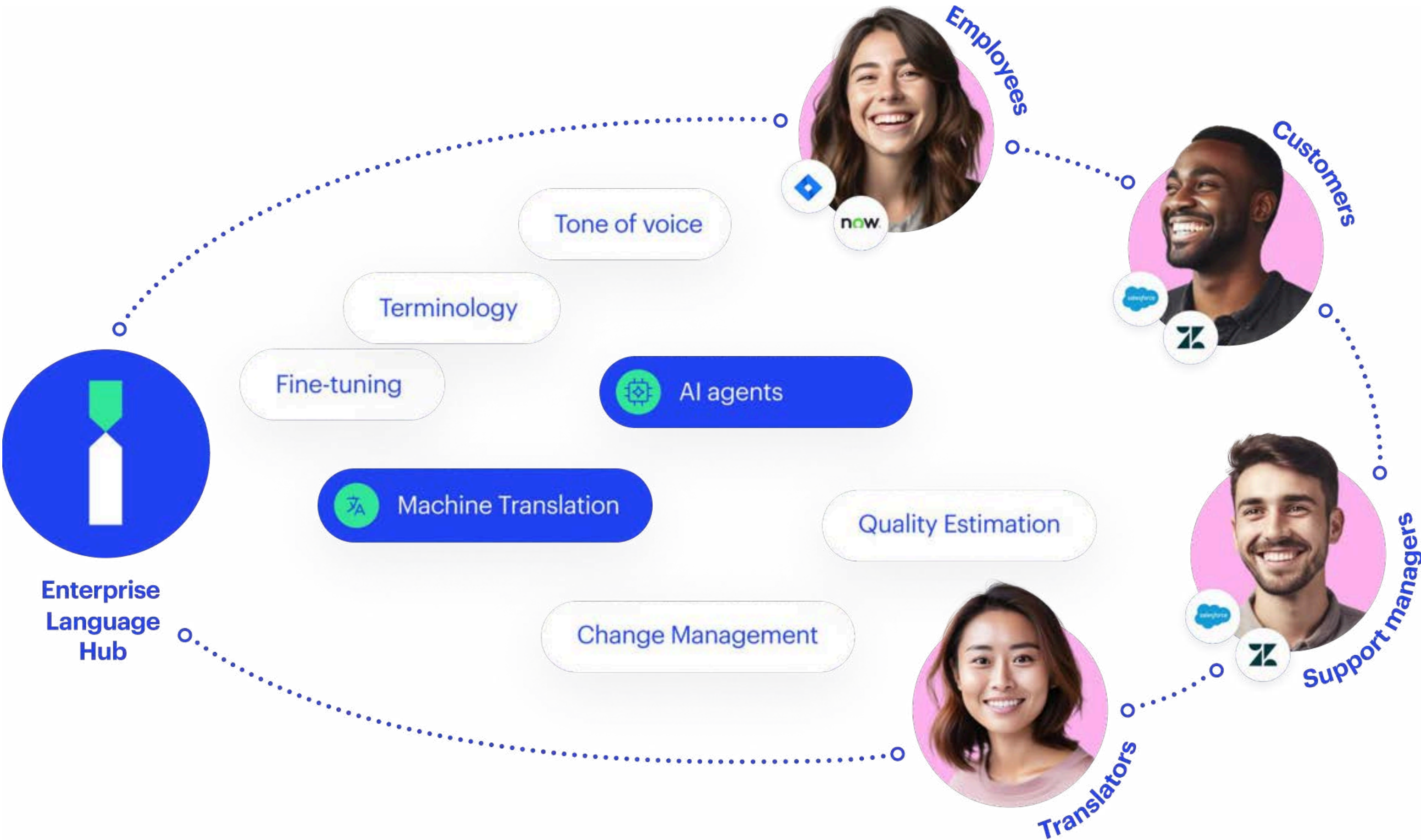
MT engines and LLMs can reach their full potential with more customization

MT and LLMs enhanced with translation memories and glossaries—plus prompt engineering and RAG for LLMs—reduce the number of errors to nearly negligible. These tools aren't optional; they're essential for professional translation quality.

Multi-Agent solutions demonstrate potential for requirement-specific translation

Analysis of Multi-Agent translation workflows showed the ability to dynamically adjust output for specific business requirements including terminology consistency, tone adaptation, and formatting requirements, suggesting a shift from general-purpose to specification-driven translation processes.

8.2 Intento – Machine Translation and Multilingual AI Agents Platform for Global Businesses



Trusted by global enterprises



8.3 Unlock Consistent and Authentic Translation Across All Touchpoints with an All-in-One, Scalable Platform

Deliver on every translation requirement

Unlike traditional MT, which requires extensive data and struggles with context, Intento Language Hub uses your translation history, glossaries, and tone preferences to deliver translations that meet your exact requirements.

Turn feedback into actionable improvements

Intento LQA provides detailed feedback comparable to human reviewers at a fraction of time and cost, identifies errors specific to your requirements, providing actionable input for automatic post-editing.

Keep existing workflows running smoothly

Intento Language Hub integrates with your localization, document management, and customer support systems so your teams keep working with tools they know and trust.

Start your enterprise AI program right from day one

We handle all aspects of implementation—from evaluating and customizing AI models to configuring multi-agent workflows and maintaining your AI program.

[Book a demo](#)

8.4 Requirements-Based Translation by Intento

Previously, we had only one option: human editing of translations. We instructed linguists on our quality expectations and relied on their expertise and available data.

Now, we can use Large Language Models as AI agents to address your specific requirements automatically. This requirements-driven translation approach uses AI Agents to **fine-tune MT with historical translations, glossary support, and tone of voice control**—even **with limited training data**. We enhance quality further through automatic quality estimation and automatic post-editing to ensure all output meets your specific quality standards.

How AI agents transform your workflow

With Intento Language Hub, you gain AI agents that work throughout the translation workflow. They **adapt, improve, and edit content both before and after translation** – correcting poorly written source content, post-editing translations based on specific expectations, checking for particular issues, and performing many more tasks to **enhance your content quality while reducing post-editing by up to 95%**.

	Data-driven translation	Requirements-based translation
Starting Point	Your historical translations	Your specific requirements
Process	Learns from past data	Built to meet defined needs
Quality guarantee	✗ "Best efforts" only	● Guaranteed compliance
Data dependency	✗ Limited by your data	● Works with minimal data
When requirements change	✗ Months to rebuild models	● Immediate adjustment
Long-term results	✗ Quality plateaus	● Consistent quality

8.5 Meet all your translation requirements with AI agents

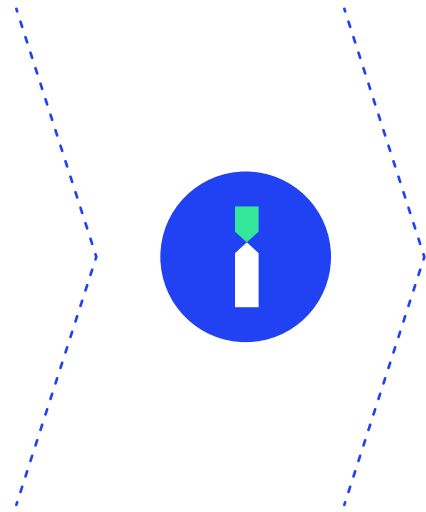
Enterprise translation quality goes beyond fluency and accuracy — it must meet specific business requirements, such as terminology, tone of voice, style guides, and so on.

Neural Machine Translation is limited to what it can learn from the training data, so with NMT you'd have to rely on humans to check translation quality, improve it if it's not good enough, and when requirements change and there's not enough new data yet.

With Intento Language Hub, you can use AI Agents instructed to check and meet all translation requirements, enabling full automation of enterprise translation.

Translation Agent

Fine-tuned on historical translations, can leverage more than 50 different NMT models and LLMs.



Post-Editor Agent

Reviews and edits translation according to instructions, can be fine-tuned on past edits

Style Guide Agent

Checks and edits translation against style guide rules

Terminology Agent

Validates against approved glossaries—flags forbidden terms and keeps terminology consistent across locales

Source Quality Improvement Agent

Improves source text before translation—fixes grammar, ambiguities, and formatting issues

Tone of Voice Agent

Adjusts translation to your brand voice without changing facts, intent, or obligations.

Compliance Agent

Checks translation to ensure that they adhere to local market regulations

8.6 **Intento LQA: Actionable, Tailored Linguistic Quality Assessment for AI-First Localization**

Intento LQA is an LLM-based, automatic quality assessment tool for AI-first enterprise localization, built that guides AI agents in delivering high-quality translations.

It provides **quick, detailed, and actionable feedback**, tailored to specific translation requirements. It highly correlates with expert judgement, while showing higher consistency than human reviewer teams.

How Intento LQA delivers quality management at scale:

Quality insights: Intento LQA provides instant quality feedback to catch errors specific to your business requirements. You focus on critical changes for faster reviews and more efficient quality control.

Quality improvement: Together with Intento AI agents, Intento LQA is used to automatically improve quality without sending it to human review.

Quality assurance: Intento LQA can be used as an additional step of the localization workflow to ensure consistent quality of the third-party content and translations with respect to specific requirements or local regulations.

[Book a demo](#)

The State of Translation Automation

An independent multi-domain
evaluation of MT engines

Commercially available
pre-trained MT models

2261 Market St, #4273
San Francisco, CA 94114

intento.to

3655 Nobel Drive, Suite 520
San Diego, CA 92122

e2f.com

Appendix A

A.1 List of Available Languages*

A.1 List of Available Languages*

Abkhazian, Achi, Achinese, Achuar-Shiwiar, Acoli, Adangme, Adhola, Afar, Afrikaans, Aguaruna, Akan, Akawaio, Albanian, Alur, Ambai, American Sign Language, Amganad Ifugao, Amharic, Anyin, Arabic, Aragonese, Argentine Sign Language, Armenian, Arpitan, Asháninka, Assamese, Asturian, Austronesian languages, Avaric, Awadhi, Ayacucho Quechua, Aymara, Azerbaijani, Baatonum, Baka (South Sudan), Balinese, Balkan Romani, Baluchi, Bambam, Bambara, Bandial, Bangla, Banjar, Bantoanon, Baoulé, Barasana-Eduria, Bari, Bariai, Baruya, Basaa, Bashkir, Basque, Batad Ifugao, Batak Karo, Batak Simalungun, Batak Toba, Belarusian, Belize Kriol English, Bemba, Berber languages, Betawi, Bhojpuri, Biak, Bikol, Bini, Bislama, Blin, Bodo, Bokobaru, Bola, Bosnian, Brazilian Sign Language, Breton, Bualkhaw Chin, Buginese, Bulgarian, Bulu, Buriat, Burmese, Busa, Cabécar, Cameroon Pidgin, Camsá, Cantonese, Catalan, Cañar Highland Quichua, Cebuano, Central Atlas Tamazight, Central Aymara, Central Bikol, Central Dusun, Central Huasteca Nahuatl, Central Kanuri, Central Kurdish, Chakma, Chamorro, Chavacano, Chechen, Cherokee, Chhattisgarhi, Chiga, Chilean Sign Language, Chinese, Chinese (Traditional), Chiru, Chokwe, Chopi, Chuukese, Chuvash, Classical Syriac, Colombian Sign Language, Coptic, Cornish, Corsican, Crimean Tatar, Croatian, Cusco Quechua, Czech, Dadibi, Danish, Dehu, Dimli (individual language), Dinka, Ditammari, Divehi, Dogri, Dombe, Duala, Dupaninan Agta, Dutch, Dyula, Dzongkha, Eastern Huasteca Nahuatl, Eastern Mari, Eastern Maroon Creole, Eastern Yiddish, Efik, English, Erzya, Esan, Esperanto, Estonian, Ewe, Falam Chin, Faroese, Fijian, Filipino, Filipino, Finnish, Finnish Sign Language, Fon, Frafra, French, Friulian, Fula, Fur, Ga, Galela, Galician, Ganda, Garifuna, Gbaya (Sudan), Georgian, German, Gilbertese, Goan Konkani, Gofa, Greek, Gronings, Guajajára, Guarani, Guerrero Amuzgo, Guerrero Nahuatl, Gujarati, Gun, Haitian Creole, Hakha Chin, Halbi, Halh Mongolian, Hausa, "Hawaii Creole English", Hawaiian, Hebrew, Herero, Hiligaynon, Hindi, Hiri Motu, Hmong, Hmong Daw, Huautla Mazatec, Huli, Hungarian, Hunsrik, Hupa, Iban, Ibanag, Icelandic, Ido, Igbo, Ika, Iloko, Inari Sami, Indonesian, Ingush, Interlingua, Interlingue, Inuktitut, Irish, Isoko, Isthmus Zapotec, Italian, Izii, Jamaican Creole English, Japanese, Javanese, Kabiye, Kabuverdianu, Kabyle, Kachin, Kakwa, Kalaallisut, Kalmyk, Kam, Kamba, Kannada, Kanuri, Kaonde, Kapingamarangi, Kaqchikel, Kara-Kalpak, Karelian, Kasem, Kashmiri, Kashubian, Kaulong, Kazakh, Kekchi, Keley-I Kallahan, Keliko, Khanty, Khasi, Khmer, Kikuyu, Kimbundu, Kinyarwanda, Kisi, Kituba (Democratic Republic of Congo), Klingon, Kok Borok, Komi, Komi-Permyak, Kongo, Korean, Krio, Kuanua, Kuanyama, Kulung (Nepal), Kurdish, Kwangali, Kyrgyz, K'iche', Ladin, Ladino, Lao, Lashi, Latgalian, Latin, Latvian, Ligurian, Limburgish, Lingala, Lingao, Lingua Franca Nova, Literary Chinese, Lithuanian, Livonian, Livvi, Lojban, Lombard, Lomwe, Low German, Lower Sorbian, Lozi, Luba-Katanga, Luba-Lulua, Ludian, Lukpa, Lule Sami, Lunda, Lundayeh, Luo (Kenya and Tanzania), Luvale, Luxembourgish, Macedonian, Machame, Madurese, Magahi, Maithili, Makasar, Malagasy, Malay, Malayalam, Male (Ethiopia), Maltese, Mam, Manipuri, Mansi, Manx, Mapuche, Maranao, Marathi, Mari, Marshallese, Marwari, Mexican Sign Language, Minangkabau, Mirandese, Mizo, Moksha, Mongolian, Morisyen, Mossi, Motu, Muscogee, Muthuvan, Māori, Nateni, Navajo, Ndonga, Neapolitan, Nepali, Newari, Ngaju, Ngäbere, Nias, Nigerian Fulfulde, Nigerian Pidgin, Niuean, Nogai, North Azerbaijani, North Ndebele, Northern Grebo, Northern Kurdish, Northern Sami, Northern Sotho, Northern Uzbek, Northwestern Ojibwa, Norwegian, Norwegian Bokmål, Norwegian Nynorsk, Nuer, Numanggang, Nyakyusa-Ngonde, Nyaneka, Nyanja, Nyankole, Nyemba, Nyungwe, Nzima, Occitan, Odia, Ojibwa, Oroko, Oromo, Ossetic, Paite Chin, Pam, Pampang, Pangasinan, Papiamentu, Paraguayan Guaraní, Parauk, Pashto, Pele-Ata, Persian, Peruvian Sign Language, Pijin, Plateau Malagasy, Pohnpeian, Polish, "Popti", "Poqomchi", Portuguese, Potawatomi, Punjabi, Quechua, Querétaro Otomi, Quíotepec Chinantec, Rarotongan, Romanian, Romansh, Roviana, Rundi, Russian, Rusyn, Ruund, Sabaot, Samberigi, Samoan, San Mateo Del Mar Huave, San Salvador Kongo, Sangir, Sango, Sanskrit, Santali, Saramaccan, Sardinian, Scots, Scottish Gaelic, Sena, Serbian, Serbian (Montenegro), Seselwa Creole French, Shan, Shipibo-Conibo, Shona, Shuar, Sichuan Yi, Sicilian, Sidamo, Silesian, Sindhi, Sinhala, Siroi, Skolt Sami, Slovak, Slovenian, Somali, Songe, South Azerbaijani, South Bolivian Quechua, South Ndebele, Southern Pashto, Southern Sami, Southern Sotho, Southwestern Dinka, Spanish, Spanish Sign Language, Sranan Tongo, Standard Latvian, Standard Malay, Suau, Sundanese, Susu, Swahili, Swati, Swedish, Syriac, Tachelhit, Tahitian, Tajik, Tamashek, Tamasheq, Tamil, Tampulma, Tatar, Tedim Chin, Telugu, Tena Lowland Quichua, Tennenet, Teso, Tetela, Tetelcingo Nahuatl, Tetum, Tetun Dili, Thai, Tibetan, Tigre, Tigrinya, Timugon Murut, Tiv, Tojolabal, Tok Pisin, Tonga (Zambia), Tongan, Tooro, Tosk Albanian, Tsimané, Tsonga, Tswa, Tswana, Tulu, Tuma-Irumu, Tumbuka, Tungag, Turkish, Turkmen, Tuvalu, Tuvinian, Twi, Tzeltal, Tzotzil, Ucayali-Yurúa Ashéninka, Udmurt, Ukrainian, Uma, Umbundu, Emoji, Northern Sotho, Upper Sorbian, Urdu, Urhobo, Uspanteco, Uyghur, Uzbek, Venda, Venetian, Venezuelan Sign Language, Veps, Vietnamese, Volapük, Vunjo, Võro, Wali (Ghana), Wallisian, Walloon, Waray, Waris, Waskia, Wayuu, Welsh, West Central Oromo, Western Bolivian Guaraní, Western Bukidnon Manobo, Western Canadian Inuktitut, Western Frisian, Western Kanjobal, Western Kayah, Western Lawa, Western Mari, Wolaytta, Wolof, Xhosa, Yabem, Yakut, Yapese, Yiddish, Yom, Yongbei Zhuang, Yongkom, Yoruba, Yucateco, Zande (individual language), Zapotec, Zarma, Zaza, Zhuang, Zotung Chin, Zulu

* in the State of Machine Translation 2024, language count has been higher due to a different processing of languages according with their ISO language codes and variants; according to the new calculations, last year there were 535 unique languages only

Appendix B

B.1 42% Correlation Showcases Both Alignment and Evaluation Challenges Between Human And Intento LQA

B.1 42% Correlation Showcases Both Alignment and Evaluation Challenges Between Human And Intento LQA

- While both Human LQA and Intento LQA use the MQM framework, they exhibit different evaluation patterns in practice.
- Human reviewers tend to flag numerous preferential issues based on personal stylistic choices and terminology preferences, while Intento LQA, as an LLM-based implementation, occasionally produces hallucinations but maintains more consistent criteria.
- LLM-based evaluation successfully identifies core quality issues like accuracy and fluency, but may not fully capture stylistic preferences and brand voice without specific instructions. By adding clear style guidelines to LLM prompts, businesses can achieve even stronger alignment with human preferences while maintaining the speed and scale benefits of automation.
- The 42% overall correlation demonstrates meaningful alignment between human judgment and Intento LQA, quite strong for translation metrics where even human reviewers often disagree. This suggests LLM-based evaluation captures important quality signals that align with human perception.
- Strong correlations in diverse language families, represented by (Ukrainian (83%), French (79%), and Korean (72%)), indicate that Intento LQA successfully adapts to different linguistic challenges.
- Lower correlations in languages like Italian may simply reflect more subjective human preferences rather than automated system limitations.

Appendix C

- C.1 Best Real-time Solutions per Language Pair Based on Human LQA
- C.2 DeepL Showcases The Best Results Among Real-Time Solutions
- C.3 Best Single-Agent Solutions per Language Pair Based on Human LQA
- C.4 OpenAI GPT-4.1 Leads Among The Single-Agent Solutions

C.1 Best Real-time Solutions per Language Pair Based on Human LQA

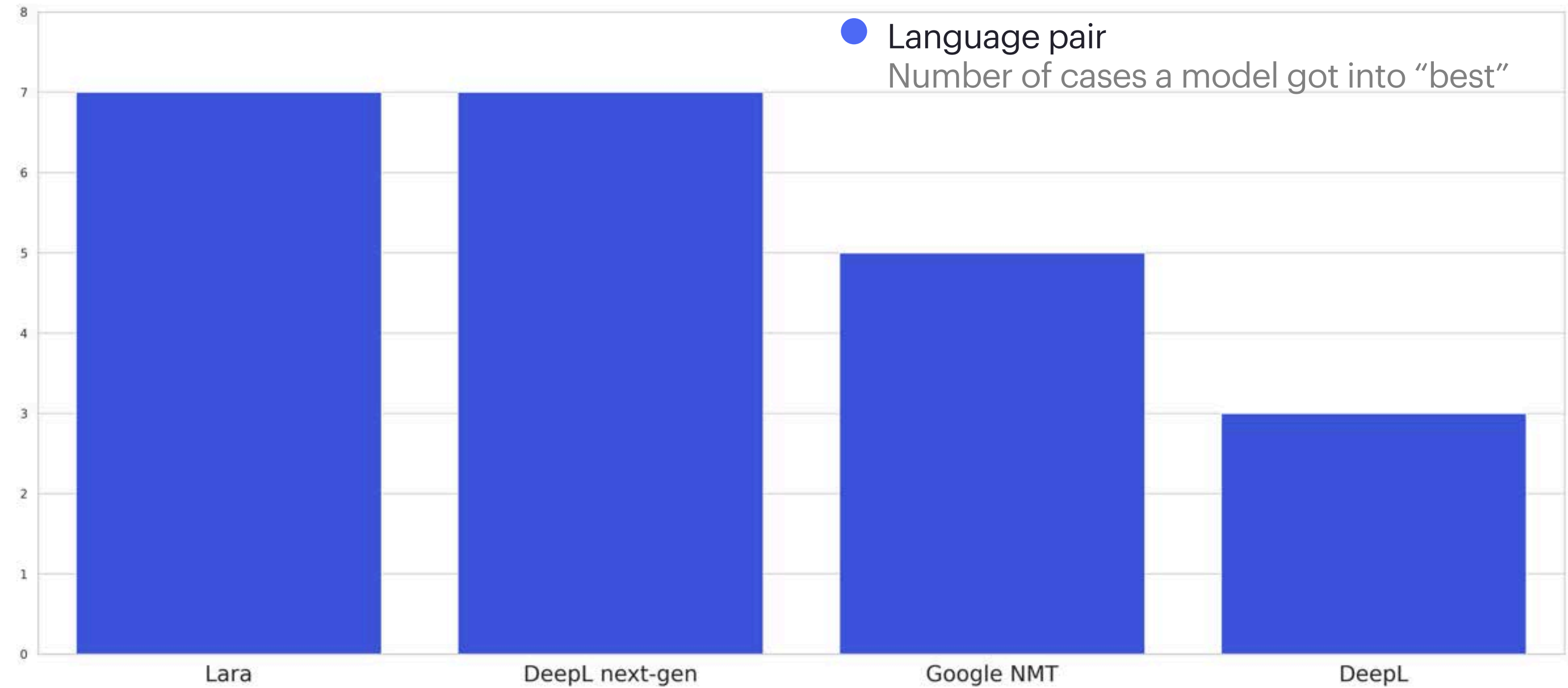
- For each language we average those scores and build a “safety band” around the mean and around the variance (83% confidence).
- A solution is only considered to be “best” if both its quality and variance intervals overlap with the highest scoring solution (see Section 3.7).

Best solutions by Human LQA			
	DeepL	Google	Translated
en-ar	Translated Lara		
en-de	DeepL	Google NMT	Translated Lara
en-es	DeepL next-gen		
en-fr	DeepL next-gen	Translated Lara	
en-it	DeepL next-gen	Google NMT	Translated Lara
en-ja	DeepL	Google NMT	Translated Lara
en-ko	DeepL next-gen		
en-nl	DeepL next-gen		
en-pt	DeepL next-gen	Google NMT	
en-uk	DeepL next-gen	Translated Lara	
en-zh	DeepL	Google NMT	Translated Lara

⚙ – Customized 📄 – Baseline

C.2 DeepL Showcases The Best Results Among Real-Time Solutions

→ [DeepL next-gen](#) and [Lara](#) demonstrate superior translation quality across multiple language pairs.



C.3 Best Single-Agent Solutions per Language Pair Based on Human LQA

- For each language we average those scores and build a “safety band” around the mean and around the variance (83% confidence).
- A solution is only considered to be “best” if both its quality and variance intervals overlap with the highest scoring solution (see Section 3.7).

Best solutions by Human LQA							
	 Anthropic	 Cohere	 DeepL	 DeepSeek	 Google	 OpenAI	 Translated
en-ar	 Cohere Command A 	 Gemini 2.5 Flash 	 Gemini 2.5 Pro 	 GPT-4.1 		 Translated Lara 	
en-de	 Claude Opus 4 	 Claude Sonnet 3.7 	 Gemini 2.5 Pro 	 Translated Lara 			
en-es	 DeepL next-gen 	 Gemini 2.5 Flash 	 OpenAI o4-mini 				
en-fr	 Google Gemma 3 	 GPT-4.1 					
en-it	 Claude Sonnet 3.7 	 Gemini 2.5 Pro 	 Google NMT 	 GPT-4.1 		 Translated Lara 	
en-ja	 Claude Opus 4 	 DeepSeek-V3 	 Google NMT 	 GPT-4.1 			
en-ko	 Claude Sonnet 3.7 	 Gemini 2.5 Pro 	 GPT-4.1 				
en-nl	 Claude Opus 4 	 DeepL next-gen 	 GPT-4.1 		 OpenAI o3 		
en-pt	 Gemini 2.5 Flash 						
en-uk	 GPT-4.1 						
en-zh	 DeepL 	 Gemini 2.0 Flash 	 Gemini 2.5 Pro 	 Google NMT 	 OpenAI o3 	 Translated Lara 	

⚙ – Customized 📄 – Baseline

C.4 OpenAI GPT-4.1 Leads Among The Single-Agent Solutions

→ OpenAI GPT-4.1 leads with 7 "best" performances, followed by Google Gemini 2.5 Pro (5) and Lara (4).

