intento e2f

The State of Machine Translation 2024

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

52 MT Engines and LLMs

11 Language Pairs

9 Content Domains

Disclaimer

March 25—May 14, 2024

The MT systems and LLMs covered in this report were accessed between March 25 and May 14, 2024. Some systems may have been updated since that time period.

Automatic scoring

This report uses semantic similarity and LLM-based DQF-MQM scores. To pick the top model for a your use case you may need a human linguist or subject matter expert review to address particular business requirements.

Stock models only

If you consider customizing an engine on your data, your choice may vary from what is suggested here. In the solutions we build for our clients, top picks often include Amazon, DeepL, Google, Microsoft, ModernMT, and Systran, depending on the languages and the amount of available training data.

Data limitations

The evaluation used plain text data. Results often differ for tagged text with some MT vendors and language pairs because of imperfect inline tag support. This report has also used segment-wise translation rather than leveraging the full text capabilities of LLMs and some MT systems.

Valid for a specific dataset

This report shows how the systems performed only on the datasets listed on slide 14. We run multiple evaluations for our clients using various language pairs and domains, and often observe different MT system rankings than those provided in this report.

There's no "best" MT system or LLM

MT performance depends on how similar your data is to the data used to train the vendors' models, their algorithms, and your quality requirements.

Trademarks

Domains? What are these?

Domain is a corpus from a specific source that may differ from other domains in topic, genre, style, level of formality et cetera*. Basically, a combination of industry sector and content type.

* as defined in "Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey" by Danielle Saunders

Executive Summary

52 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	

9 Content domains

General	Entertainment	Healthcare
Colloquial	Hospitality	Legal
IT	Education	Financial

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

20 MT engines and LLMs show the best results for some language pairs and domains (55% are LLMs).

- Amazon
- Baidu
- Claude 3 Haiku
- Claude 3 Opus
- Claude 3 Sonnet
- Command R+
- DeepL
- GPT-4
- GPT-4 Turbo
- GPT-4o
- Gemini Pro 1.5
- Google
- Microsoft
- ModernMT
- NiuTrans
- PaLM 2 Chat Bison
- PaLM 2 Text Bison
- PaLM 2 Text Unicorn
- Tarjama
- Yandex

Language expansion across all MT engines

191K unique language pairs

+1K compared to 2023

658 unique languages

The machine translation landscape is evolving as multilingual LLMs capable of high-quality translations are developed and deployed at an accelerating pace.

We've evaluated 52 systems overall (+15 to 2023), among which there are 24 Large Language Models (+19 to 2023).

55% of all top-performing models are LLMs (it was 17% in 2023). The largest LLM is not always the best for translation (even from the same provider).

In 25% of all cases, LLMs are significantly better than any MT. More in Colloquial, Education and Entertainment. In 12% cases, MT is better than any LLM. More in English to Arabic, and in IT and Legal domains.

LLMs are 10-100 times less expensive than MT systems, but 50-1000 times slower, so we provide a separate rating for real-time systems. On average, it comes at 11% penalty in quality.

Among analyzed pairs and domains, Colloquial domain and English-Arabic pair carry the most critical issues due to complexity of translation.

Large Language Models for Translation

1. Expansion across the board

The Large Language Model market has experienced explosive growth in recent years. Among 52 models we have assessed in this report, 24, nearly half, are Large Language Models.

2. Large Language Models are in the 1st tier

Large Language Models, such as GPT-4o, PaLM2 Text Unicorn, and Gemini Pro 1.5, demonstrate performance comparable to top-tier commercial MT systems across most language pairs. While their cost is 10 to 100 times lower, LLMs have a latency 50 to 1,000 times higher than traditional MT engines.

3. LLMs are priced 10 to 100 times lower than traditional MT

On average, LLMs are priced 10 to 100 times lower than traditional MT engines, making them a highly attractive alternative for companies looking to reduce costs without compromising on quality for human post-diting scenarios.

4. 50-1000 slower than traditional MT

Although LLMs offer lower costs compared to traditional MT engines, their translation times are typically 50 to 1,000 times slower, rendering them unsuitable for real-time translation applications.

5. Open-source LLMs are generally in the 2nd tier

While the performance of open-source LLMs like TowerInstruct 7B v0.2 or Command R approaches top-tier commercial engines, the majority of open-source LLMs produce lower-quality translations due to their more limited multilingual capabilities compared to their commercial counterparts.

6. Customization is possible

The performance of LLMs can be enhanced through the use of Retrieval-Augmented Generation (RAG) or prompt engineering techniques. These methods allow for adjustments in tone of voice, mitigation of gender bias, and incorporation of domain-specific terminology. Moreover, several LLMs can be fine-tuned for translation tasks by leveraging existing translation memories.

About Intento

Intento is a machine translation and multilingual generative AI platform for global enterprise companies.

Our Enterprise Language Hub enables companies like Procore and Subway to deliver consistent, authentic language experiences across all markets and audiences. It combines machine translation and generative AI models into automatic translation workflows, customizing them to client data and integrating them into customers' existing software systems for localization, marketing, customer support, and other business functions.

With Intento, clients achieve high-quality, real-time translations for all users and team members worldwide. The Enterprise Language Hub is ISO-27001 certified, ensuring enterprise top-tier security for GenAI solutions in high-demand industries. Intento also offers ISO-9001-certified expert help for setting up and maintaining MT and AI models and constantly refines these models with new data and user feedback.

Trusted by the global enterprise

PROCORE Government of Canada vmware Agilent playrix

We have been evaluating stock Machine Translation models since May 2017. For customers, we also evaluate customizable NMT models (you can get a glimpse here).

As we show in this report, the Machine Translation landscape is complex and dynamic. Models from five different vendors are required to achieve the best quality in popular language pairs, with a dramatic price difference (as much as 200 times.)

Book a demo

Enterprise Language Hub

Machine Translation and multilingual Generative AI platform for global businesses. We deliver immediate, tailored, and personalized language experience in all the software systems your customers and teammates already use, supporting over 650 languages.

Book a demo

Language Hub for

Localization

Save up to 95% on translations by combining best-fit Machine Translation with source quality improvement and automatic post-editing based on GenAI. Let us take care of your custom MT engines. Works with over 15 TMS.

lokalise memoq RWS XTM International ...

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.

COGNIGY salesforce VERINT zendesk ...

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 Atlassian, Microsoft, ServiceNow, Zendesk, and other tools.

ATLASSIAN chrome metricstream Microsoft ...

Enterprise GenAI Portal

Boost your team's productivity with GenAI-based automatic language skills (tone of voice, summarization, and more) while keeping your security team happy. Choose the right GenAI model for every task and amplify it with Machine Translation for non-English content. Pay for actual usage, not seats.

ANTHROP\C Google Microsoft OpenAI

About e2f

Established in 2004, e2f helps people and machines understand each other fluently, regardless of language, content, and culture. e2f solutions empower Fortune 50 brands to monitor, objectively assess, and improve communications on a global scale.

e2f delivers world-class translation and training data with its proprietary technology stack for translation, quality review, and AI services. e2f offers a global resource pool of skilled professionals in virtually all countries and languages.

To learn more, contact e2f or visit website.

e2f services

- → MT detection and MT quality evaluation services that enable organizations to monitor suppliers for compliance with brand standards for human and machine translation.
- → Creation of custom Lingosets™, or augmented multilingual datasets that represent real human conversational flow. Lingosets serve as benchmarks for conversational AI deployments.
- → Golden datasets and training datasets that enable leading MT providers to evaluate and fine-tune engine performance.

Overview

1. MT Engines
2. Datasets
3. Evaluation Methodology
4. Evaluation Results
5. Miscellaneous
6. Takeaways

52 Machine Translation Engines and Large Language Models

11 Language Pairs

9 Content Domains

1. MT Engines and Large Language Models

1.1 Machine Translation Landscape

Generic stock models

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

Vertical stock models

- Alibaba
- Baidu
- CloudTranslation
- Lingua Custodia
- Microsoft
- NiuTrans
- Omniscien
- PROMT
- RoyalFlush
- SAP
- SYSTRAN (by CHAPSVISION)
- Tartu NLP
- Tarjama
- Ubiqus
- XL8

Custom terminology support

- Amazon
- Baidu
- DeepL
- Google
- IBM
- Microsoft
- Rozetta
- RWS
- SYSTRAN
- Ubiqus
- Yandex

Static domain adaptation

- Alibaba
- AppTek
- Baidu
- CloudTranslation
- Globalese
- Google
- IBM
- KantanAI
- Microsoft
- Omniscien
- PangeaMT
- Prompsit
- SYSTRAN (by CHAPSVISION)
- Tilde
- Ubiqus
- PROMT
- Rozetta
- RWS
- Yandex

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.

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.

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.

Vertical Stock Models

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

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 re-training.

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

Baseline only

- **Aleph Alpha** — Luminous
- **Anthropic** — Claude 3[4] (Opus, Sonnet, Haiku), 2.1, 2.0, Instant
- **Google** — PaLM2 Unicorn-001, Gemini 1.0 Ultra, Gemini 1.5 Pro
- **Microsoft Azure** — OpenAI GPT-4, GPT-4 Turbo, Mistral Large, other open models
- **OpenAI** — GPT-4 Turbo[3]

Cloud Commercial

Customizable

- **AI21** — Jurassic-2 Ultra[4], Mid[4], Light
- **Amazon AWS** — Titan, Cohere Command, Meta Llama 2
- **Cohere** — Command[4]
- **Google** — PaLM 2 Bison, Gemini 1.0 Pro
- **Microsoft Azure** — OpenAI GPT-3.5
- **Mistral** — Small[3,4], Large[3,4]
- **NVIDIA** — Nemotron-3[3]
- **OpenAI** — GPT-3.5 Turbo[3], GPT-4[5]

Open Commercial[1]

- **01.AI** — Yi
- **Ai21** — Jamba
- **Alibaba** — PolyLM, Qwen, Qwen1.5[2]
- **AllenAI** — OLMo
- **BAAI** — Aquila2
- **Baichuan** — Baichuan 2
- **BAIR** — OpenLLaMA, Starling
- **Big Science** — BLOOM[2]
- **Cerebras** — Cerebras-GPT
- **Cohere** — Aya, Command
- **Databricks** — Dolly 2.0, MPT, DBRX[3]
- **Deci** — DeciLM[3]
- **Eleuther AI** — GPT-Neo, GPT-NeoX, Pythia, Polyglot
- **Google** — T5-FLAN, Gemma
- **HuggingFace** — Zephyr
- **Lianjiia Tech** — BELLE
- **LLM360** — Amber, Crystal
- **LLMZoo** — Phoenix
- **Meta AI** — Llama 2[2,3,4], Llama 3[2,3,4]
- **Microsoft** — Phi, Phi-1.5, Phi-2, Phi-3
- **Mistral** — Mistral 7B[3,4], Mixtral 8x7B[3,4]
- **Preferred Networks** — PLaMO
- **Salesforce** — XGen
- **Silo AI** — Poro, Viking
- **Snowflake** — Arctic
- **Stability AI** — Stable LM 2
- **StatNLP** — TinyLLaMA
- **TII (UAE)** — Falcon[3]
- **X.ai** — Grok-1

Open Non-commercial

- **BAAI** — Aquila2 70B
- **BAIR** — Koala
- **Meta AI** — Llama
- **Stability AI** — Stable LM Zephyr, Stable Beluga 1/2
- **Stanford** — Aplaca, Vicuna
- **Unbabel** — TowerInstruct

[1] Apache 2.0 or MIT licenses
[2] limited commercial use, read license terms for details
[3] available as a cloud commercial model via Azure
[4] available as a cloud commercial model via AWS
[5] experimental

Updated on April 24, 2024. For any revisions, please reach out to us at hello@inten.to.

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

1.3 Evaluated MT Engines and LLMs

Customization options

- None
- TM
- Glossary
- Both

MT Engines

Alibaba Cloud General	Alibaba E-Commerce MT	Amazon Translate	Baidu Translate API	DeepL APi	Elia MT API	HiThink RoyalFlash Finance Translation
Globalese Machine Translation	Google Cloud Advanced Translation	Kawamura by NICT Translation Engine	Lingua Custodia Machine Translation API	Microsoft Language Translator	Mirai Translator	ModernMT Adaptive
Naver Papago NMT Commercial	NiuTrans Translation Cloud Platform	Oracle Machine Translation	Pangeanic Machine Translation API	PROMT Machine Translation	SYSTRAN PNMT	Tarjama MT API
TartuNLP Neurotõlge MT	Tencent Cloud TMT API	Tilde Machine Translation API	TREBE Machine Translation API	Ubiqus Translation API	Yandex Translate API	Youdao Cloud Translation API

LLMs

Aya-101 Cohere	Claude 3 Haiku Anthropic	Claude 3 Opus Anthropic	Claude 3 Sonnet Anthropic	Command-R Cohere	Command-R+ Cohere	Gemini Pro Google Vertex
Gemini Pro 1.5 Google Vertex	GPT-3.5 Turbo OpenAI	GPT-4 OpenAI	GPT-4o OpenAI	GPT-4 Turbo OpenAI	Jurassic Ultra AI21	LLaMA-2 Meta AI
LLaMA-3 Meta AI	Mistral Large Mistral AI	Mixtral 8x7B Mistral AI	PaLM2 Chat Bison Google Vertex	PaLM2 Text Bison Google Vertex	PaLM2 Unicorn Google Vertex	RakutenAI 7B Instruct RakutenAI
Titan Text Express Amazon AWS	TowerInstruct 13B v0.1 Unbabel	TowerInstruct 7B Internal v0.2 Unbabel				

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

2. Datasets

2.1 Preparation

The source data collection and initial cleaning were done by Intento.

Open-Source English Texts

Carefully selected from open-source data

→ Found several resources for each domain and selected the ones with suitable license agreements

→ Extracted high-quality segments

Data samples for various domains are used according to their licence agreements: Financial data, Hospitality data 1, Hospitality data 2, Legal data, Entertainment data, IT data, Colloquial data

Filtering to Ensure High-Quality Source

Collected data for 9 domains using open-source resources

→ Removed duplicates, tags, and broken symbols

→ Removed segments under 4 words

→ Removed segments that were truncated (except for the Colloquial sector) and segments that were longer than one sentence

→ Manually checked each segment in every domain to avoid segments with an ambigous meaning or incorrect tone of voice

2.1 Preparation

The dataset translations and quality assurance were done by e2f.

Translation by Native Speaking Experts

→ Selected native translators with expert-level qualifications and positive feedback in each language and domain.

→ For reviews, selected native language experts in editing and proofreading across multiple domains, and positive customer feedback.

→ Proofread strings supplied by Intento for compliance with proper English grammar, spelling, and punctuation and supplied files to translators via e2f's Translation, Editing, and Proofreading (TEP) platform.

Quality Assurance

Provided via e2f's TEP portal

→ Human translations were compared with ones generated by the leading machine translation engines using e2f's MT Detection tool, and determined the probability that they contained machine-translated and/or post-edited content (MTPE).

→ Strings whose MTPE probability exceeded e2f's threshold triggered expert review and was followed by re-translations, which were automatically reassessed. The resulting golden dataset does not bear traces of MTPE.

→ Quality assurance reports were run on capitalization, punctuation, spelling, numbers, spaces, and typos. Reviewers implemented necessary changes and proofread the dataset prior to final delivery.

2.2 Content Domains and Language Pairs

9

content domains
per language pair

11

language pairs
per domain

	Colloquial	Education	Entertainment	Financial	General	Healthcare	Hospitality	IT	Legal
en-ar	467	471	467	472	473	480	470	473	476
en-de	472	475	471	472	473	481	473	475	474
en-es	469	474	466	472	471	478	466	470	469
en-fr	473	472	473	475	473	475	472	471	473
en-it	474	476	471	475	472	478	478	476	473
en-ja	477	475	472	471	474	473	477	473	473
en-ko	472	478	465	469	470	471	470	470	474
en-nl	473	470	466	472	472	470	476	475	475
en-pt	472	472	472	472	476	481	476	471	472
en-uk	473	474	470	475	475	473	474	476	479
en-zh	474	472	469	476	475	478	474	473	473

2.3 Content Samples by Domain

General

“Walmart is also the largest grocery retailer in the United States.”

Finance

“Both operating profit and net sales for the three-month period increased, respectively from €16m and €139m, as compared to the corresponding quarter in 2006.”

Hospitality

“Very reasonably priced and the food is excellent, I had pasta which was delicious, and my friend had the Italian meats & cheeses.”

Healthcare

“Leishmaniosis caused by Leishmania infantum is a parasitic disease of people and animals transmitted by sand fly vectors.”

Legal

“Landlord and Tenant acknowledge and agree that the terms of this Amendment and the Existing Lease are confidential and constitute proprietary information of Landlord and Tenant.”

Entertainment

“Further, they are aided by a magnificent cast of co-stars, most notably their secretary, played by Isabel Tuengerthal, who is a rare gem with great comic potential.”

Education

“Find what straight lines are represented by the following equation and determine the angles between them.”

IT

“The interface is in Python, a dynamic programming language, which is very appropriate for fast development, but the algorithms are implemented in C++ and are tuned for speed.”

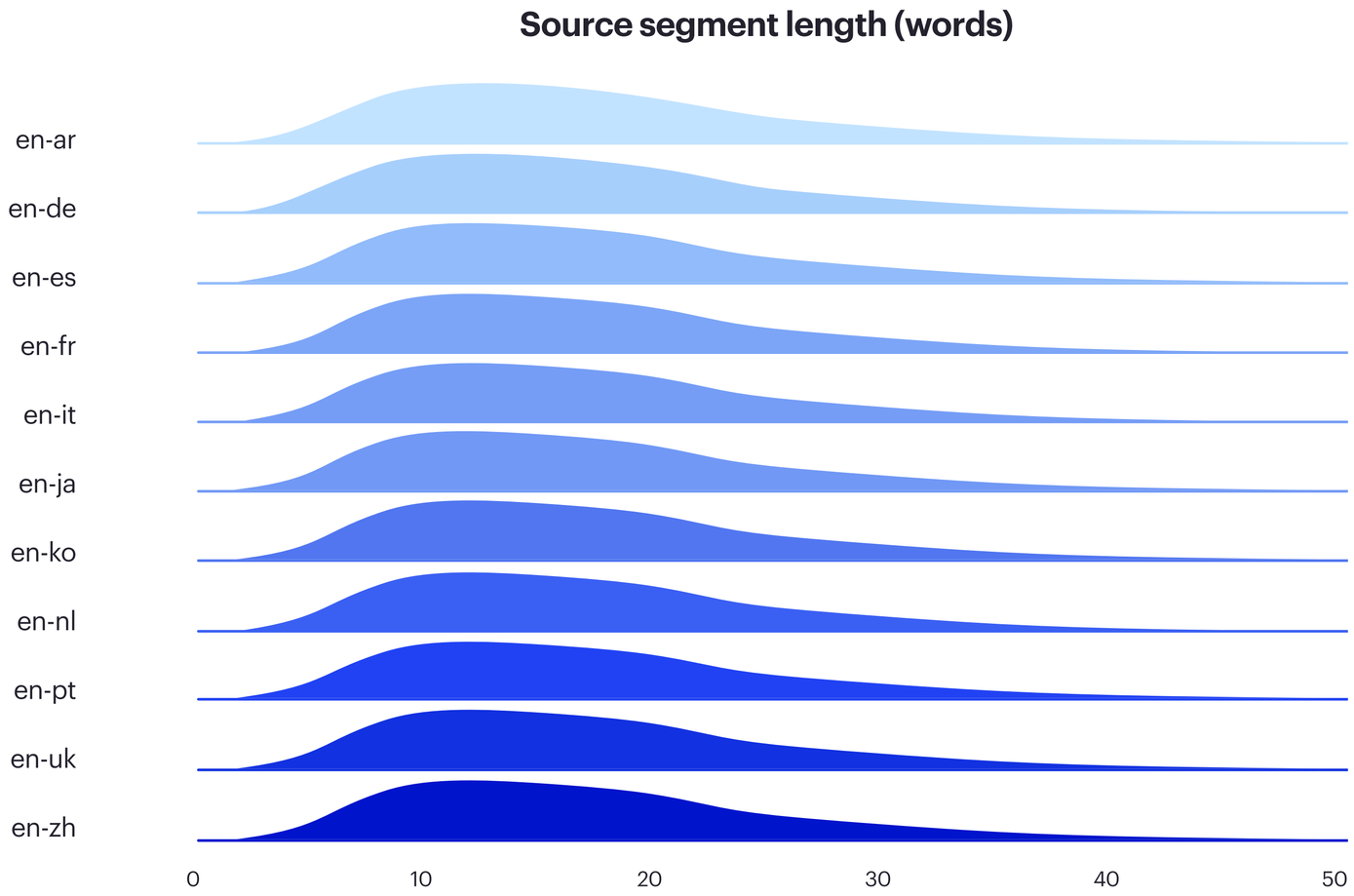
Colloquial

“and, in fact, there are two huge lenses that frame the figure on either side”

2.4 Sentence Length

→ 11 pairs with English source were translated in total

→ Sentences that were too short (< 4 words) were excluded from the dataset.



3. Evaluation Methodology

3.1 Scores to Choose From

hLEPOR

Syntactic similarity

Compares similarity of token-based n-grams. Penalizes both omissions and additions. Penalizes paraphrases / synonyms. Penalizes translations of different length.

paper + code

TER

Syntactic similarity

Measures the number of edits (insertions, deletions, shifts, and substitutions) required to transform a machine translation into the reference translation. Penalizes paraphrases/ synonyms. Penalizes translations of different length.

paper + code

SacreBLEU

Syntactic similarity

Compares token-based similarity of the MT output with the reference segment and averages it over the whole corpus. Penalizes omissions and additions. Penalizes paraphrases / synonyms. Penalizes translations of different length.

paper + code

BERTScore

Semantic similarity

Analyzes cosine distances between BERT representations of machine translation and human reference (semantic similarity). Does not penalize paraphrases / synonyms. May be unreliable for terminology in domains and languages underrepresented in BERT model.

paper + code

☆ COMET

Semantic similarity

Predicts machine translation quality using information from both the source input and the reference translation. Achieves state-of-the-art levels of correlation with human judgement. May penalize paraphrases/synonyms. The version of the model used in the report is wmt22-comet-da.

paper + code

☆ Intento LQA

Large Language Model-based

Analyses the quality of machine translation based on both source and reference translation using DQF-MQM framework. Achieves high correlation with human assessment. May penalize paraphrases/synonyms.

3.2 Choosing Engines for Linguistic Quality Assessment using COMET

→ While COMET provides valuable insights into the performance of various models, it is important to note that relying solely on COMET may not give a comprehensive understanding of the models' strengths and weaknesses.

→ In several combinations of domain x pair, up to 24 models fall within the 83% confidence interval, which mean further analysis is necessary to fully assess the nuances and differences between these models beyond COMET scoring.

→ For this evaluation, we also used Intento LQA metric based on the DQF-MQM framework for machine translation evaluation.

→ We have first scored the translations using COMET to get the ranking of providers for each combination of language pair x domain.

→ We have then chosen top-runners for each combination based on the following conditions:

- We chose top-runners from the the 90th percentile of COMET score in each combination of domain x pair
- We made sure that there were no less than 3 providers per language pair x domain
- If there were only LLMs, we added a low-latency model
- If more than one model from one LLM family appears, unless it’s a provider with reasonable tiers between models, we removed all but one with the highest COMET

3.3 Issue Classification and Severity for Intento LQA

We use the following issue classification as given in the DQF-MQM framework when working on Intento LQA:

- Accuracy issues: Addition, Omission, Mistranslation, Over-translation, Under-translation, Untranslated text
- Fluency issues: Punctuation, Spelling, Grammar, Grammatical register, Inconsistency, Link/cross-reference, Character encoding
- Terminology issues: Inconsistent use of terminology
- Style issues: Awkward, Inconsistent style, Unidiomatic
- Design issues: Length, Local formatting, Markup, Missing text, Truncation/text expansion
- Locale convention issues: Address, Date, Currency, Measurement, Shortcut key, Telephone format
- Verity issues: Culture-specific reference
- Other issues

As seen in DQF-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

3.4 Examples of Issue Classification Using Intento LQA

Intento LQA score: 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 أتمنى لو كان بإمكاني أن أقول شيئًا أفضل عن Take The High Ground أحب طاقمها الموهوب، ومخرجها الموهوب ريتشارد بروكس، وحتى الموضوع السخيف لديمتري تيومكين ونيد واشنطن، الجديد من جائزة الأوسكار قبل عام عن High Noon. ”

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.

Intento LQA score: 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。”

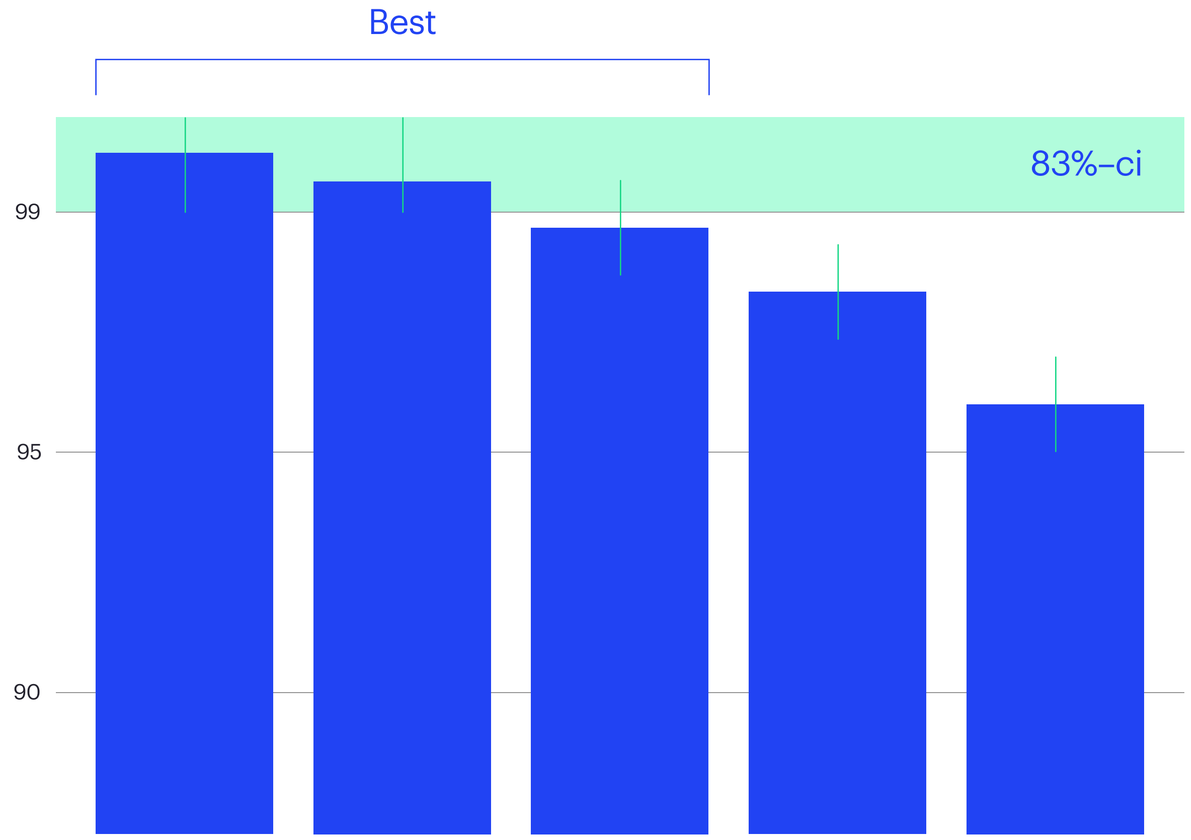
Intento LQA:
- Major error (5-point deduction) - Addition: The machine translation adds '欧州' (European) which is not present in the source text.
- Major error (5-point deduction) - Mistranslation: The machine translation uses 'つながった' (led to) which implies a past action, whereas the source text implies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

3.5 How We Choose Best Engines Using Intento LQA

1. Score all segments according to their penalties, where 100 is a maximum score obtainable by an MT engine or an LLM, and 100-Σ(all segment penalty points) is the final segment score

2. Average segment-level scores across the corpus

2. Identify a group of top-runners (BEST) within a an 83% confidence interval[1,2] of the leader



1. Harvey Goldstein; Michael J. R. Healy. The Graphical Presentation of a Collection of Means, Journal of the Royal Statistical Society, Vol. 158, No. 1. (1995), p. 175-177.

2. 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.

4. Evaluation Results

4.1 Best MT Engines per Domain

→ In the next two slides, we show the best MT engines by Intento LQA score. Each square shows the best providers for a particular language pair in a specific domain. The color of the square shows the achievable MT quality for this domain compared to other domains in this language pair.

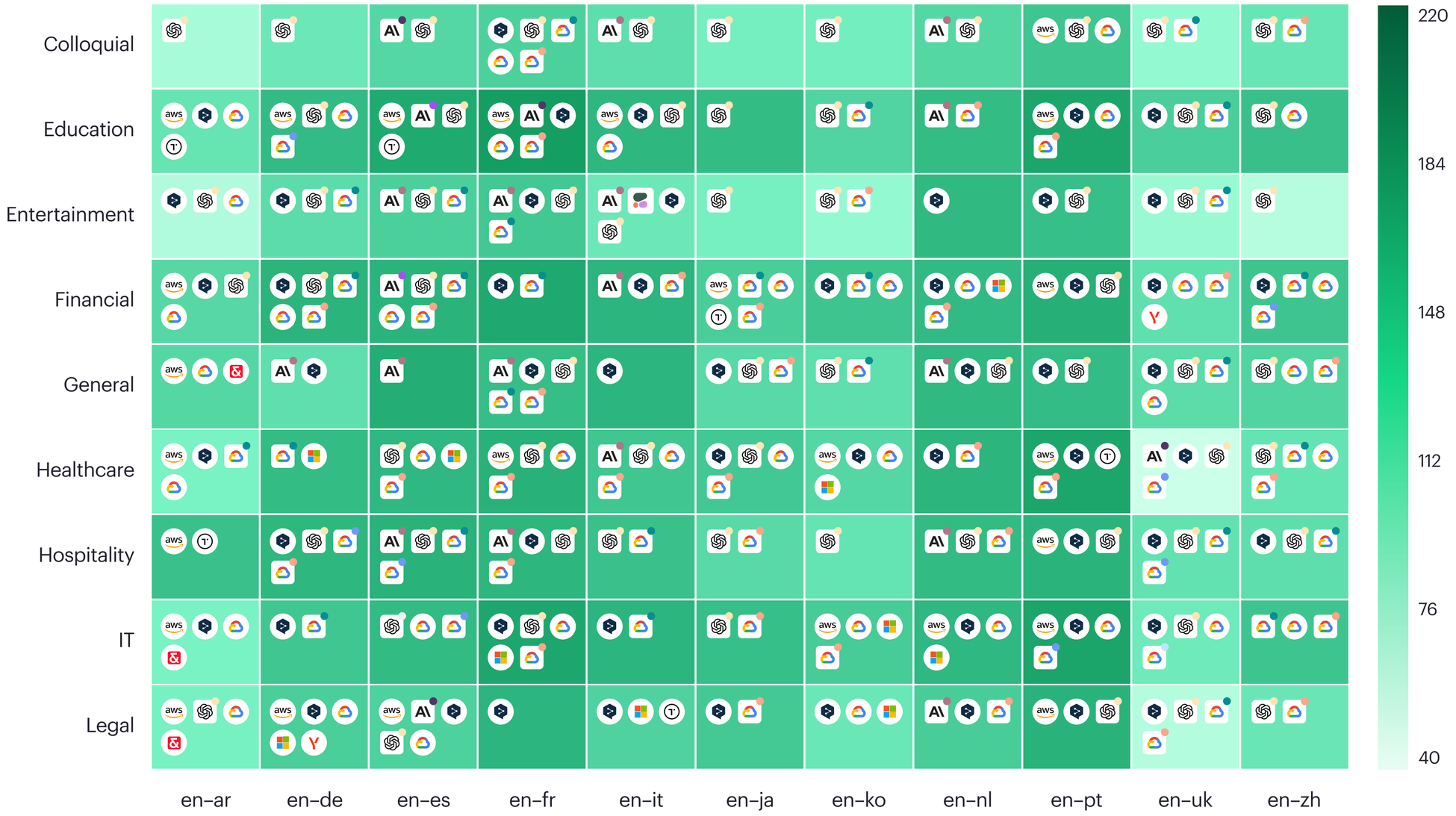
→ For example, we see that the best engine for the English-Japanese pair in the Educational and Entertainment domains is GPT-4o. Its score for the Educational domain is higher, so we expect less post-editing than in the Entertainment domain.

→ We showcase both the best MT engines and LLMs overall and also, separately, the best engines suitable for real-time translation, as LLMs have high latency and might not be the best choice for such a scenario.

→ The score values are not comparable between different language pairs.

→ Engines in one bucket provide the best quality for this language pair and domain, with no statistically significant difference between them. They are presented in alphabetical order.

Available quality and best commercial MT engines and LLMs by domain per Intento LQA score

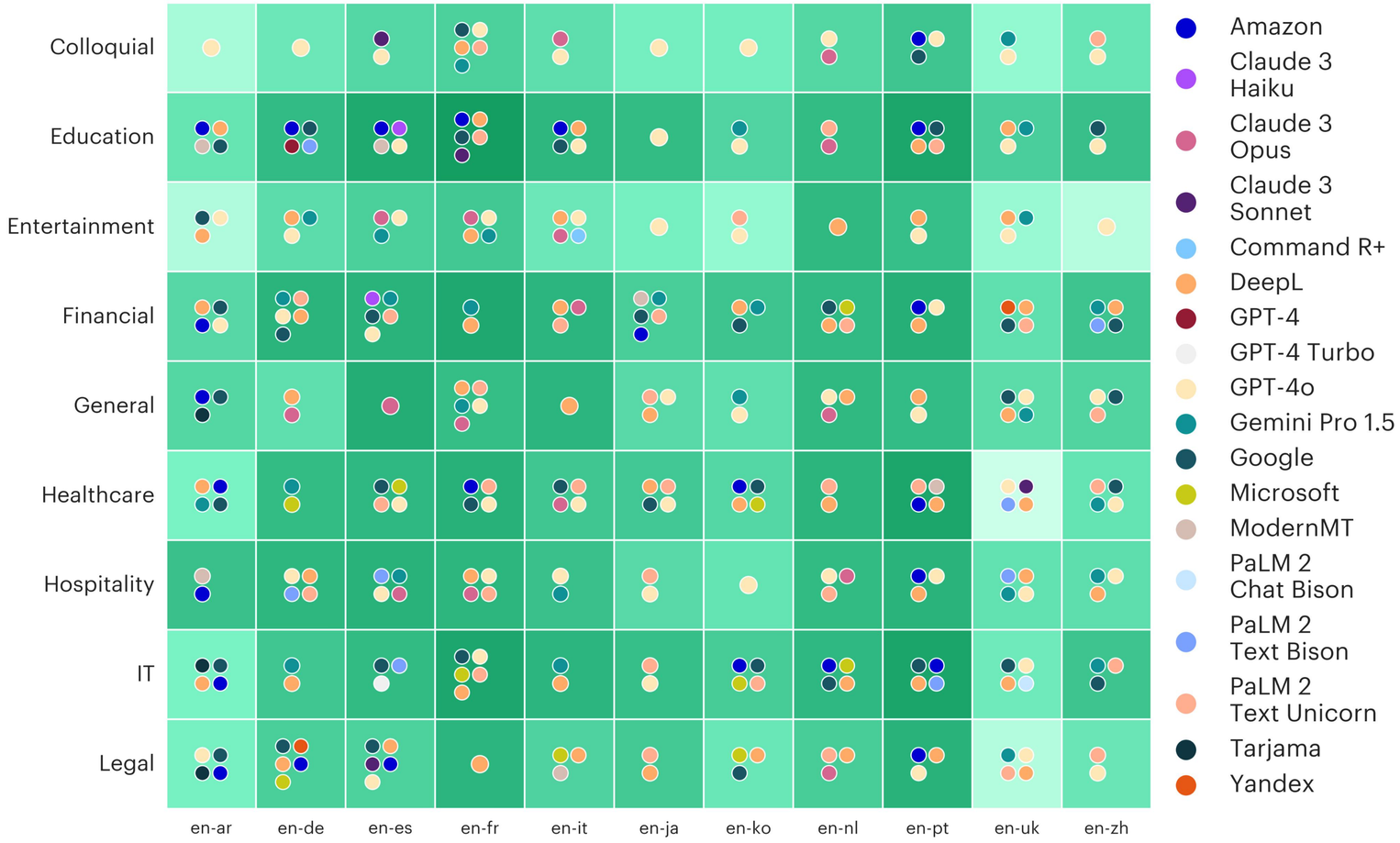


Engines are shows in alphabetical order as they are statistically non-distinguishible and are in the same tier.

4.2 Eighteen Providers Show the Best Results

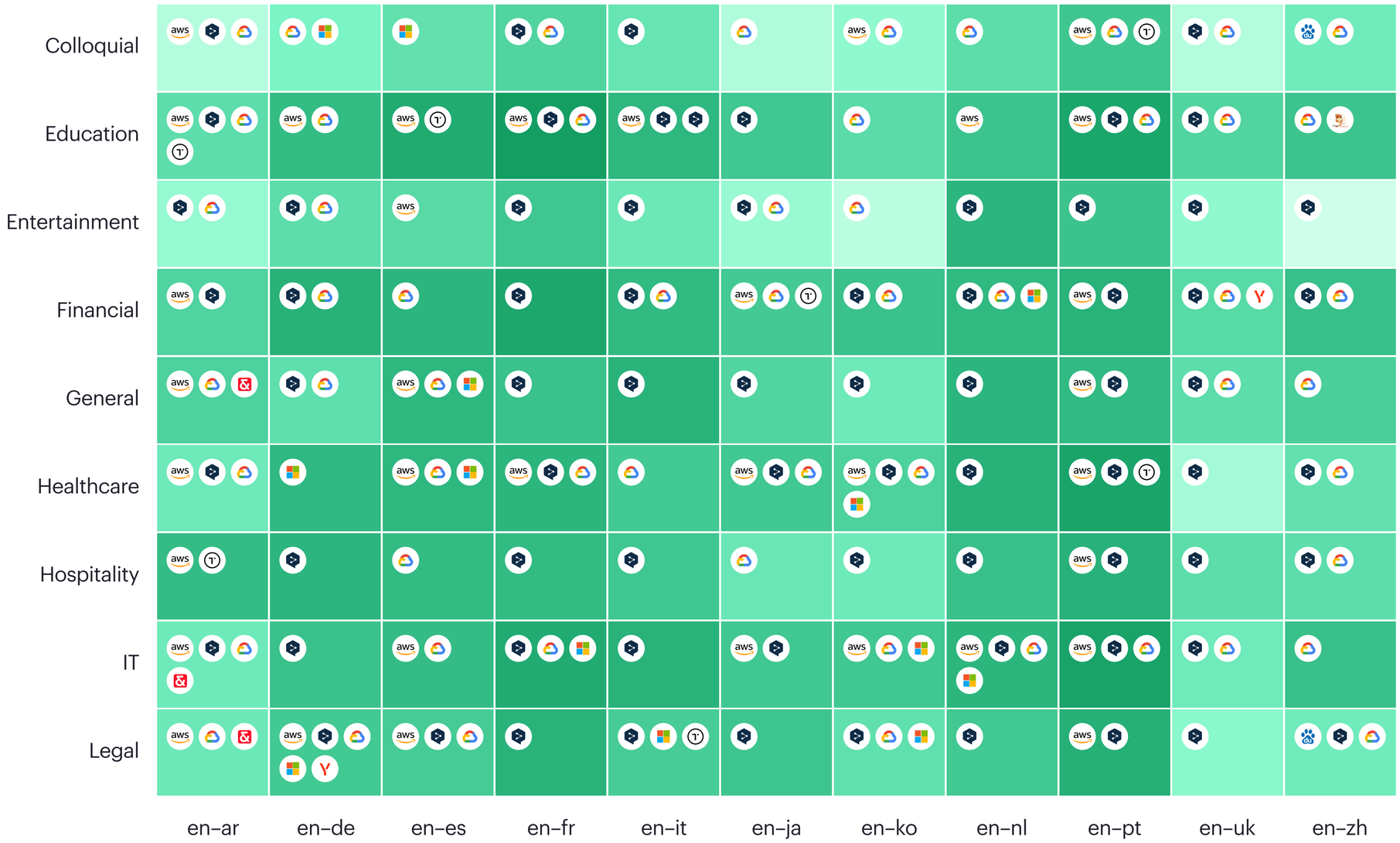
- → Many engines perform best with English to Spanish, Portuguese, and French.
- → Colloquial and Entertainment domains, as well as Japanese and Korean languages, require a careful choice of MT vendor, as relatively few perform at the top level.
- → Despite having several comparable engines per language pair, Colloquial and Entertainment domains show relatively low scores, which may indicate the importance of customization or context.
- → Recently published GPT-4o outperforms several traditional MT engines across multiple domains and pairs.

Available quality and best commercial MT engines by domain per Intento LQA score



Engines are shows in alphabetical order as they are statistically non-distinguishible and are in the same tier.

Available quality and best commercial MT engines by domain per Intento LQA score (real-time translation)*



- Amazon
- Baidu
- DeepL
- Google
- Microsoft
- ModernMT
- NiuTrans
- Tarjama
- Yandex

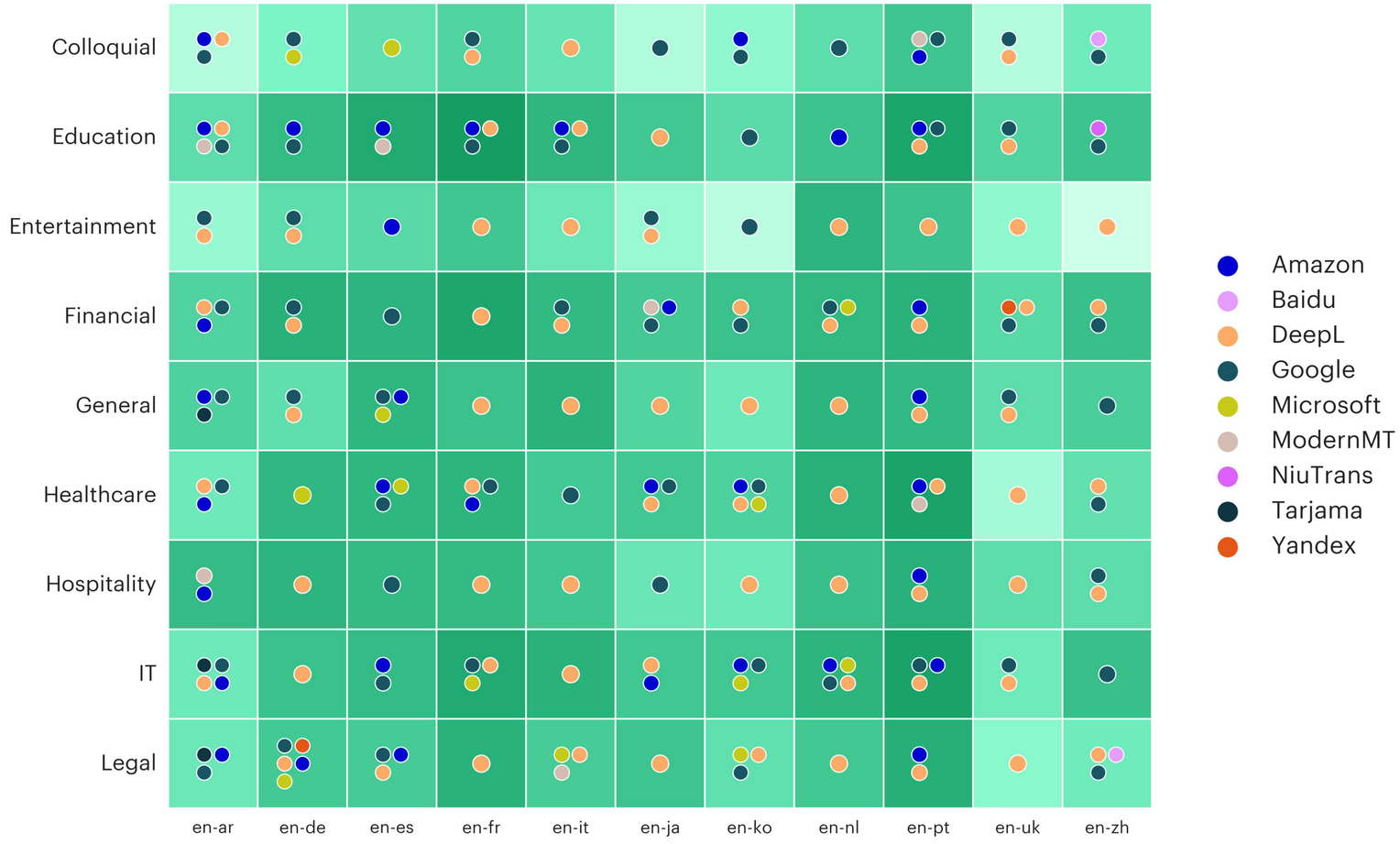
* When working with real-time translation engines only, decrease in quality should be expected compared to LLM providers, which is represented as percentages in the bottom right corner of each bucket

** Engines are shows in alphabetical order as they are statistically non-distinguishible and are in the same tier.

4.3 Nine MT Engines Excel Among Real-Time Providers

- → When it comes to real-time translation, several engines perform best with English to Spanish, Arabic, and Dutch.
- → Arabic, Korean and Ukrainian languages require a careful choice of MT vendor when working with real-time translation as few perform at top-level.
- → Entertainment and Colloquial domains show relatively low scores, which highlights the importance of customization and context.
- → Google and DeepL showcase superior translation performance in multiple domain x pair combinations.

Available quality and best commercial MT engines by domain per Intento LQA score (real-time translation)*



When working with real-time translation engines only, an average 1% decrease in quality should be expected compared to LLM providers

Engines are shows in alphabetical order as they are statistically non-distinguishible and are in the same tier.

4.4 Six Engines Provide Minimal Coverage

6 MT engines and LLMs provide minimal coverage* for all pairs and industries, 1-4 per domain.

Education
Amazon, Claude 3 Opus, GPT-4o

General
Amazon, Claude 3 Opus, DeepL, GPT-4o

Financial
DeepL, Gemini Pro 1.5

IT
DeepL, GPT-4o, Google

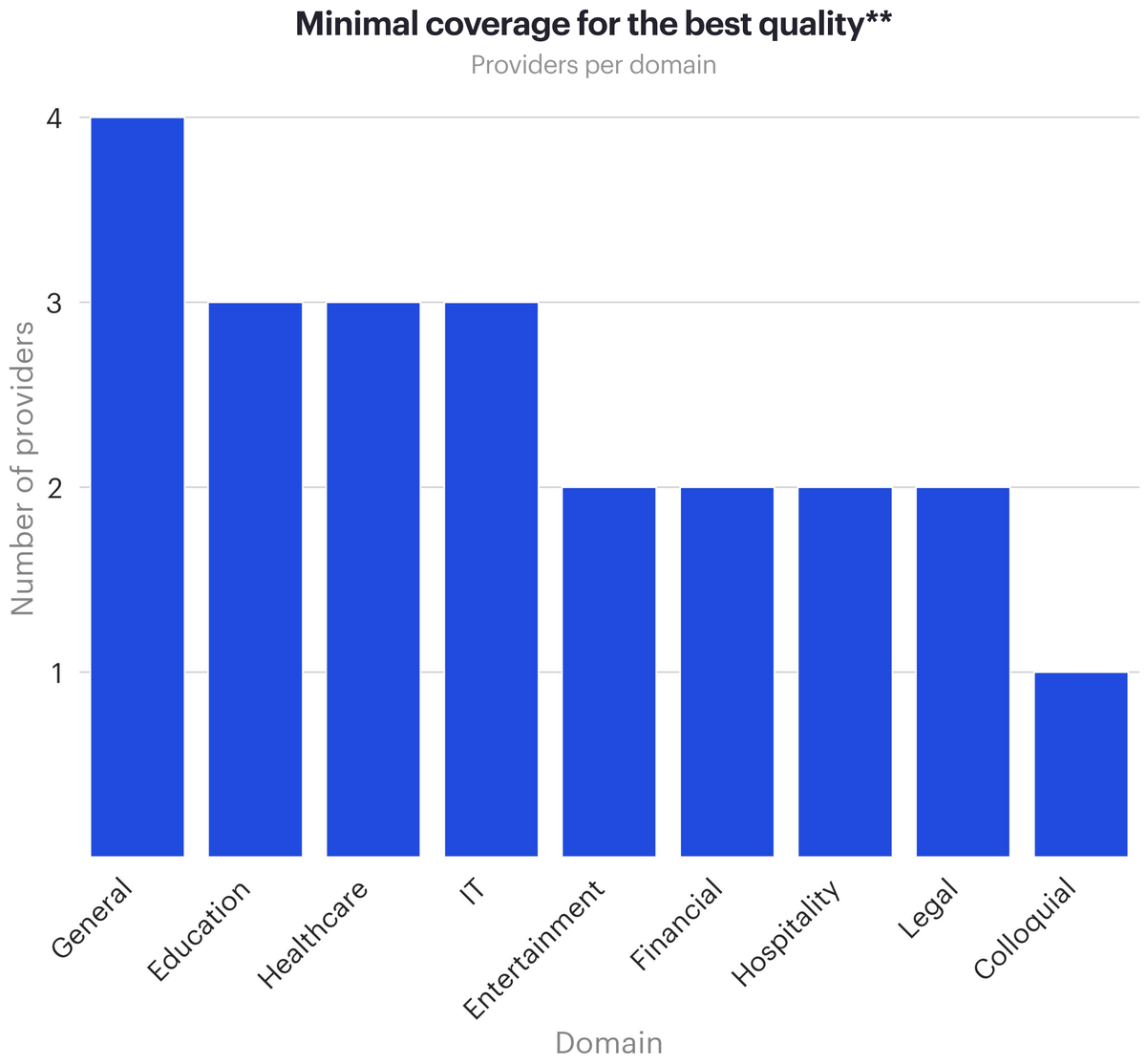
Healthcare
DeepL, Gemini Pro 1.5, Google

Entertainment
DeepL, GPT-4o

Legal
DeepL, GPT-4o

Hospitality
Amazon, GPT-4o

Colloquial
GPT-4o



* For every domain, we provide the minimum number of providers needed to translate all language pairs in this specific domain.

** Engines are shows in alphabetical order as they are statistically non-distinguishible and are in the same tier.

4.5 GPT-4 and DeepL Consistently Outperform Other Models

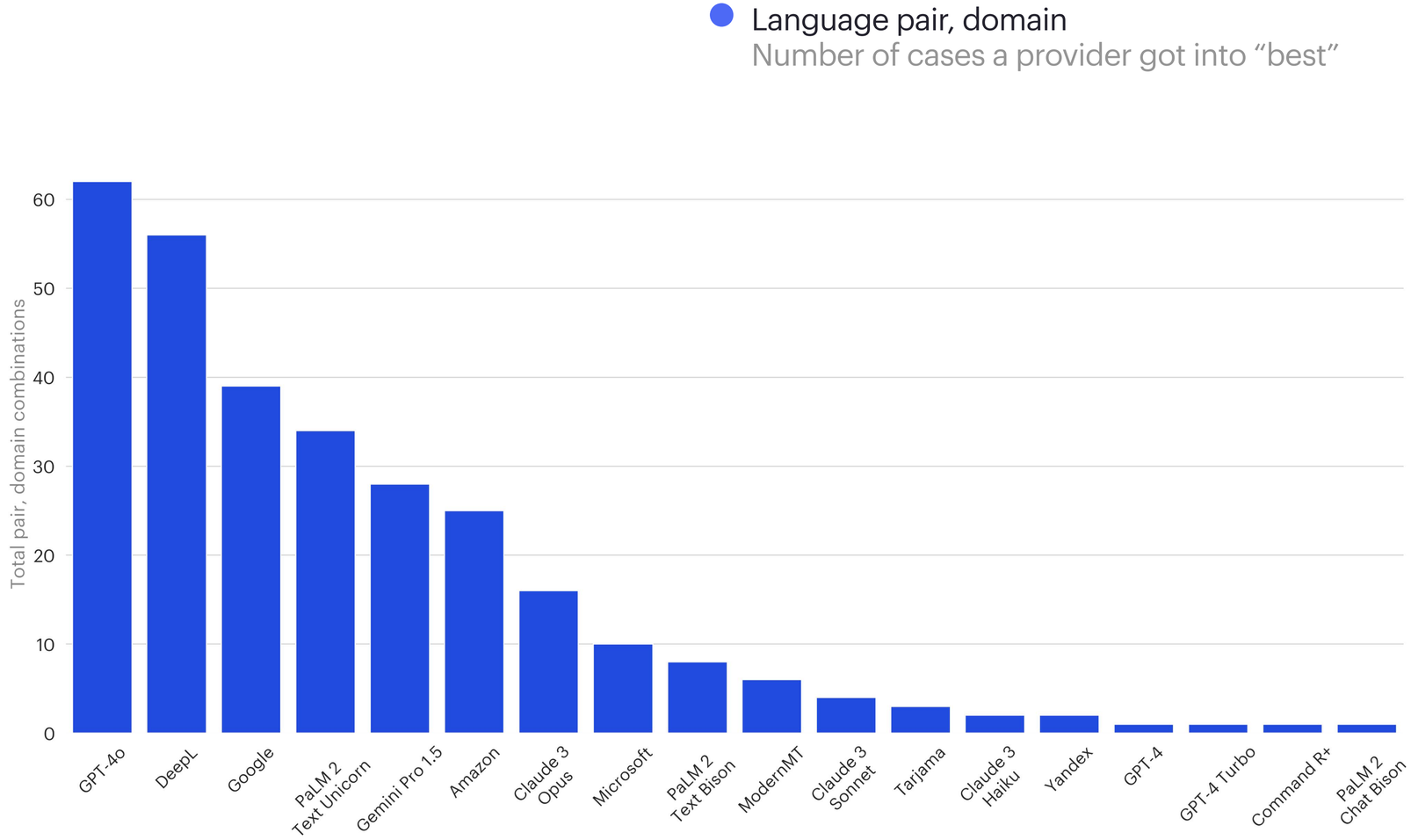
11 language pairs, 9 domains

Some providers were tested only in their specific domains and language pairs:

→ HiThink RoyalFlush specializes in en-zh translation in the Finance domain

→ TREBE specializes in Iberian languages, and was used for en-es translation

→ Tarjama specialized in Arabic translation



4.6 Four Real-Time Engines Provide Minimal Coverage

4 MT engines provide minimal coverage* for all pairs and industries, 2–3 per domain, in the real-time scenario.

Colloquial
DeepL, Google, Microsoft

Education
Amazon, DeepL, Google

Entertainment
Amazon, DeepL, Google

Hospitality
Amazon, DeepL, Google

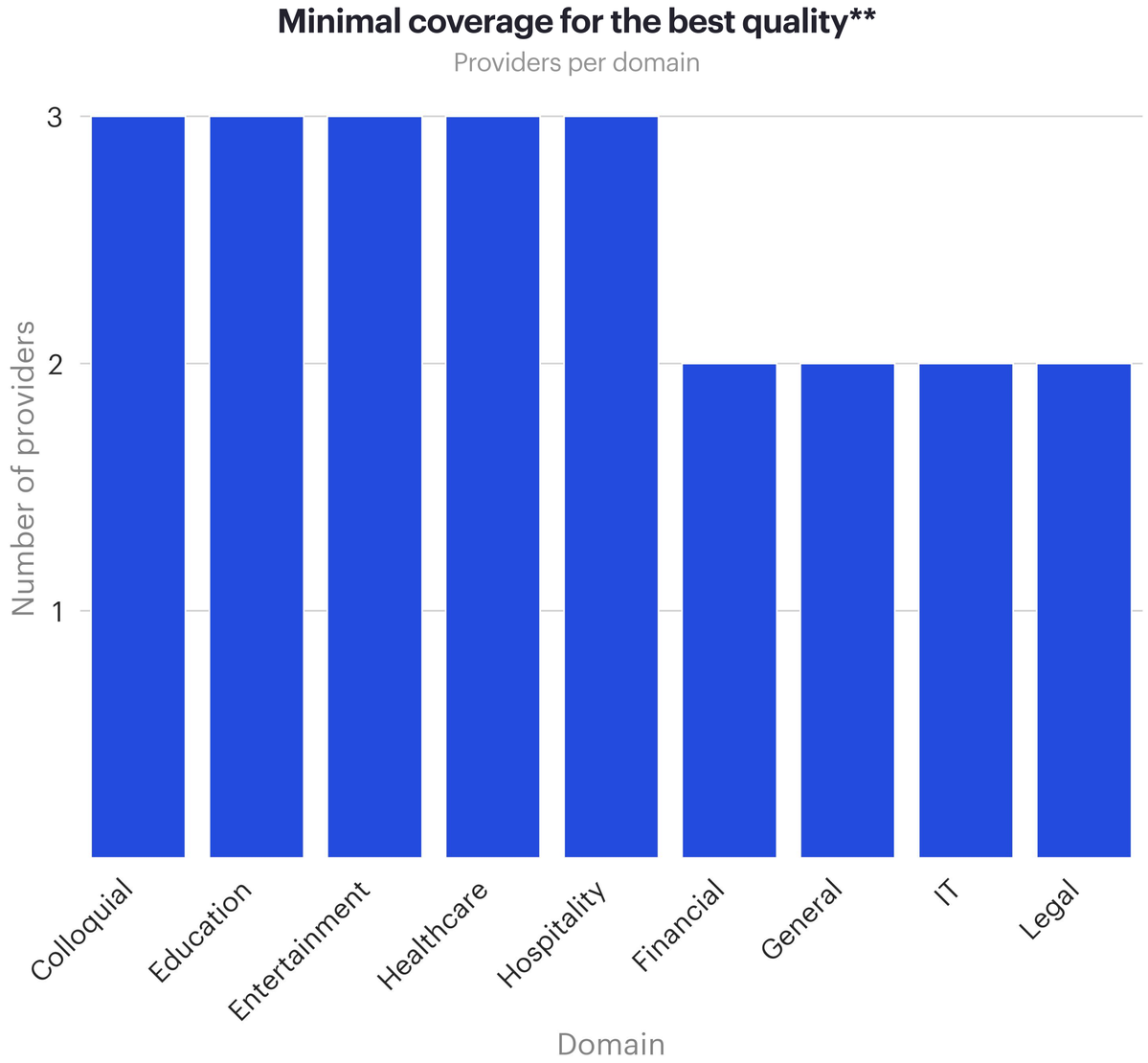
Healthcare
DeepL, Google, Microsoft

Legal
Amazon, DeepL

Financial
DeepL, Google

IT
DeepL, Google

General
DeepL, Google



* For every domain, we provide the minimum number of providers needed to translate all language pairs in this specific domain.

** Engines are shows in alphabetical order as they are statistically non-distinguishible and are in the same tier.

4.7 DeepL Surpasses Other Real-Time Engines

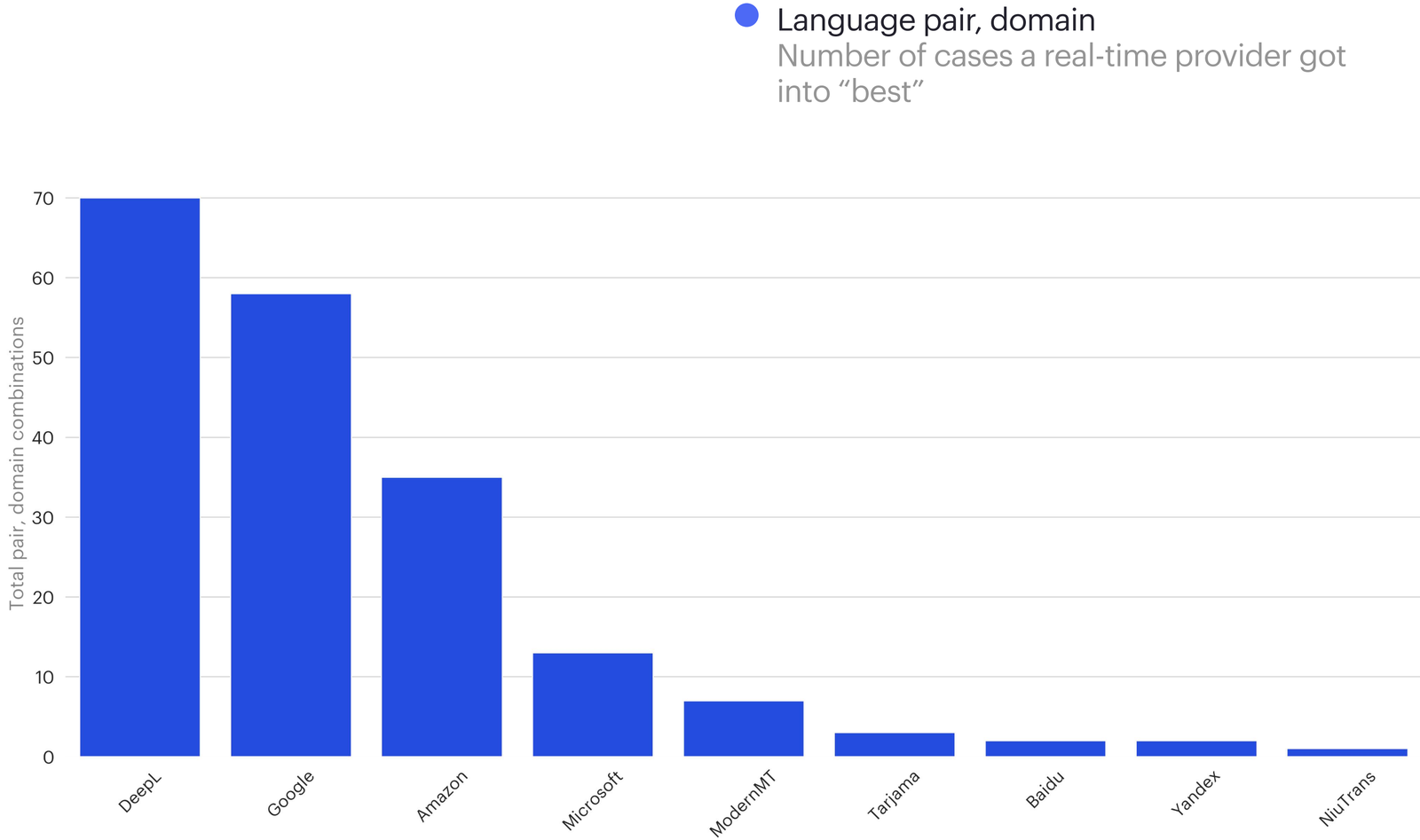
11 language pairs, 9 domains

Some providers were tested only in their specific domains and language pairs:

→ HiThink RoyalFlush specializes in en-zh translation in the Finance domain

→ TREBE specializes in Iberian languages, and was used for en-es translation

→ Tarjama specialized in Arabic translation



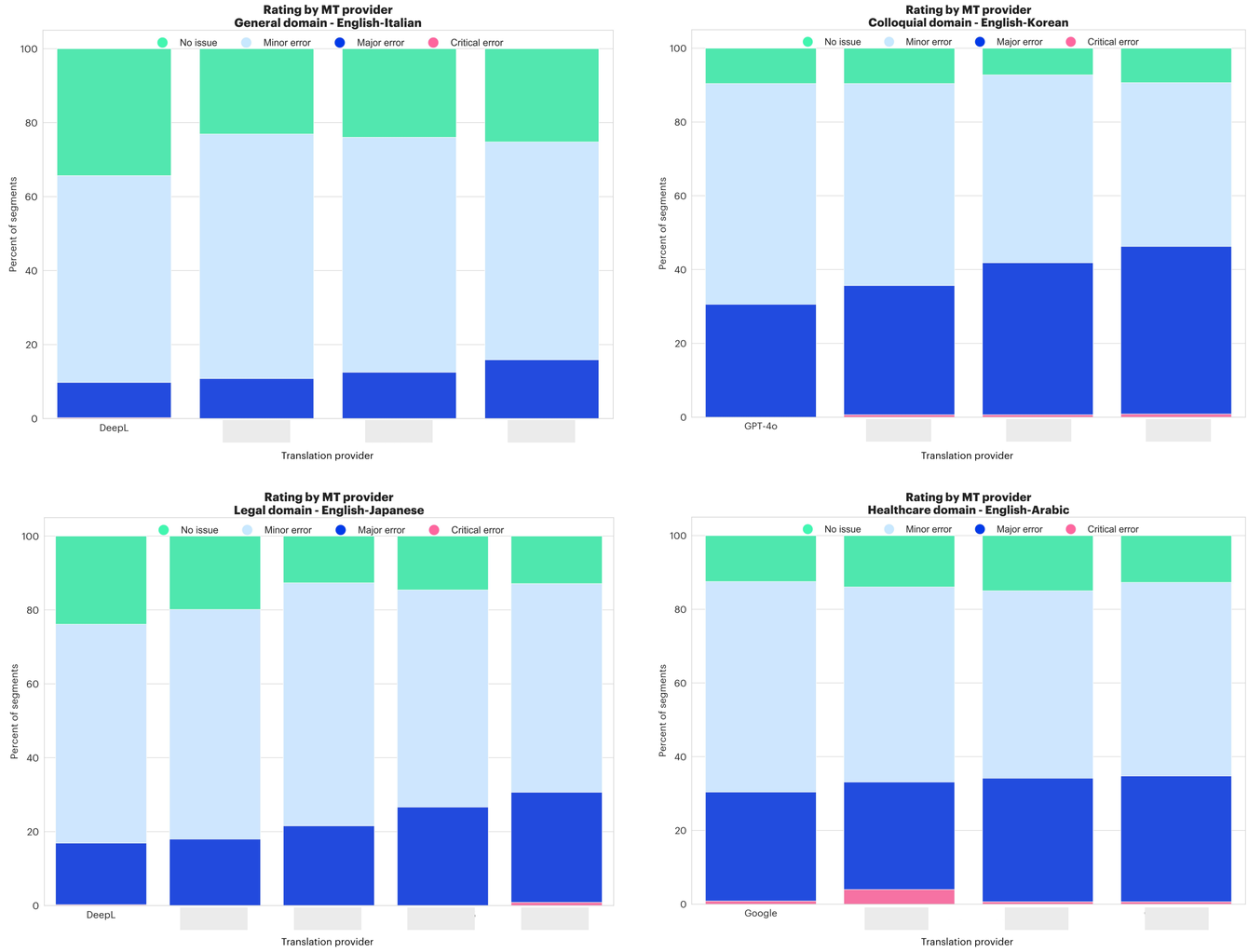
4.8 Relatively Low Number of Major or Critical Errors

→ Across all combinations domain × pair, GPT-4o, DeepL, and Google have the most segments with no or minor issues.

→ Among analyzed pairs and domains, Colloquial domain and English-Arabic pair carry the most major and critical issues due to complexity of translation.

→ According to the DQF-MQM framework, minor issues are described as having a 'slight impact on meaning.' This broad definition leads to a large proportion of segments being classified as having minor issues.

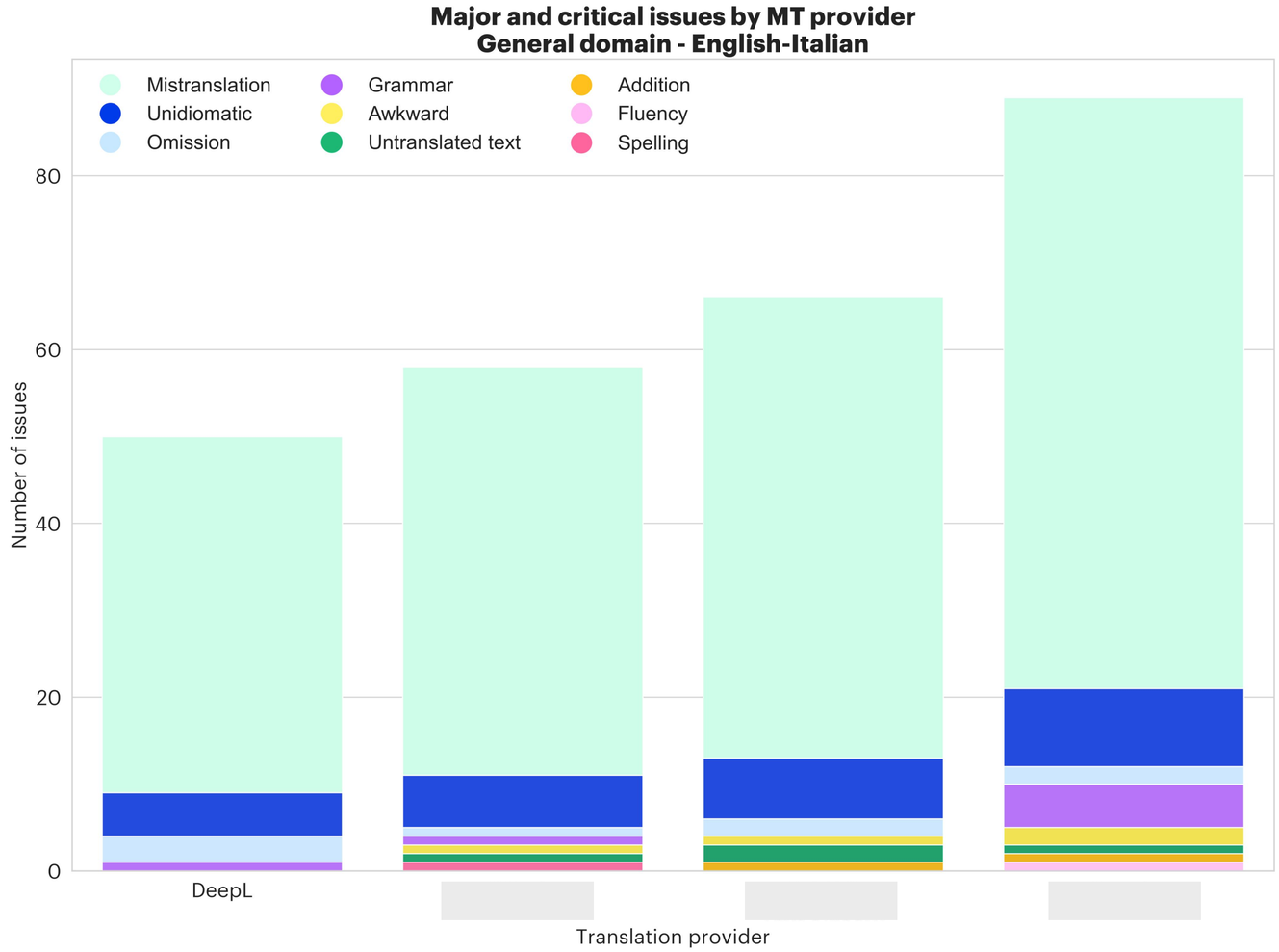
→ Higher linguistic quality can be achieved using engine customization and glossary support.



We present an example of ratings in one combination of domain x pair to showcase general distribution between different translation issues and lack thereof

4.9 Mistranslations Are More Common than Other Translation Issues

- → Mistranslations represent 80% of all major and critical issues.
- → Most major mistranslation showcase translations with altered overall meaning, because of incorrectly used words or phrases.
- → Unidiomatic translations tend to appear where a non-literal translation is possible but MT did not handle it properly.
- → The rest of the issues overall represent only 10% of the data which proves that MT keeps on drastically improving every year.
- → We present examples of major issues on the next slide.



* Major or critical issues present in one segment are counted separately

** We present an example of ratings in one combination of domain x pair to showcase general distribution between different translation issues and lack thereof

4.10 Examples of Major and Critical Issues

Mistranslation

Source: “Terribly undercooked pasta, not sure if they have even heard the term al dente pasta as it was hardly cooked at all.”

MT: “パスタがひどく茹ですぎで、アルデンテという言葉を知っているのか疑問に思うほどでした。”

Under-translation

Source: “Florence Rice runs the gamut from comedienne to heroine.”

MT: “Florence Rice passe de la comédie à l'héroïne.”

Grammar

Source: “Of course the mozzarella is astounding, but the bread and meats and everything else are also just fantastic.”

MT: “Natürlich ist die Mozzarella erstaunlich, aber auch das Brot, die Fleischwaren und alles andere sind einfach fantastisch.”

Unidiomatic

Source: “"Pardon My Pups" is an enjoyable little film, with Shirley Temple stealing all her scenes as the hero's lively kid sister.”

MT: “Вибачте моїх цуценят — це приємний невеликий фільм, у якому Ширлі Темпл викрадає всі свої сцени як жива сестра героя.”

Omission

Source: “You have to have a strong stomach and a firm grip on yourself to sit through this, and I wouldn't recommend trying unless you have a good reason.”

MT: “除非有充分的理由，否则我不建议您尝试。”

Untranslated text

Source: “The yellow to red color of many cheeses, such as Red Leicester, is normally formed from adding annatto.”

MT: “عادة ما يتشكل اللون الأصفر إلى الأحمر للعديد من أنواع الجبن، مثل Red Leicester، من إضافة annatto.”

We present an example of ratings in one combination of domain x pair to showcase general distribution between different translation issues and lack thereof

5. Miscellaneous

5.1 191,010 Language Pairs Across All MT Engines*

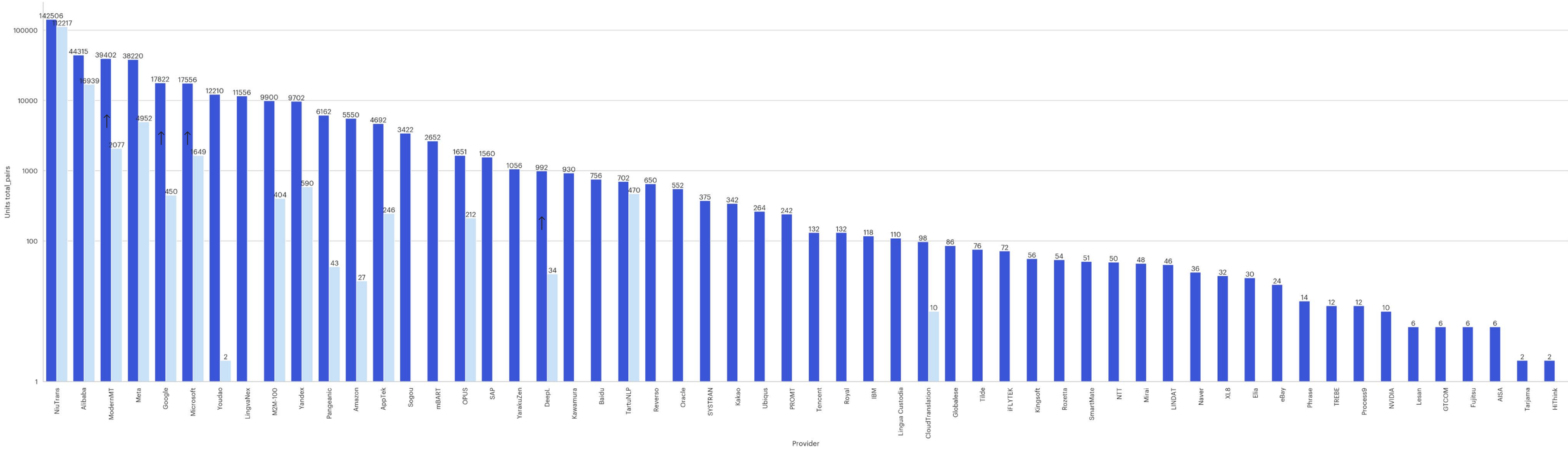
- total language pairs
- unique language pairs
- ↑ language pair growth

From 190,085 in May'23 to 191,010 in May'24

Several new languages added by ModernMT, Google, Yandex, Microsoft, and DeepL

Yakut, Emoji, Uzbek Cyrillic, and Indian language Bodo, are new to Intento

Added new niche MT provider Tarjama specializing in Arabic translations



* 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).

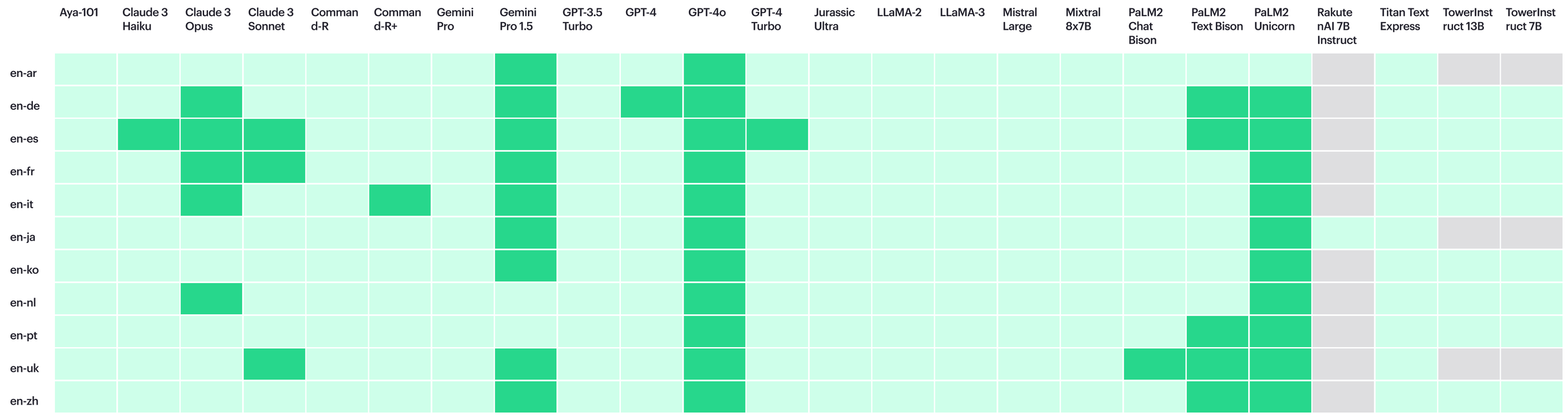
*** Due to LLMs multilinguality and the nature of the data they were trained on, there is no definitive list of languages or language combinations they support. For now, we do not include LLMs in this slide.

5.2 LLM Multilinguality Does Not Mean Equal Support of All Languages

pair is supported

pair is supported and the model appeared in the list of best engines

pair is not supported

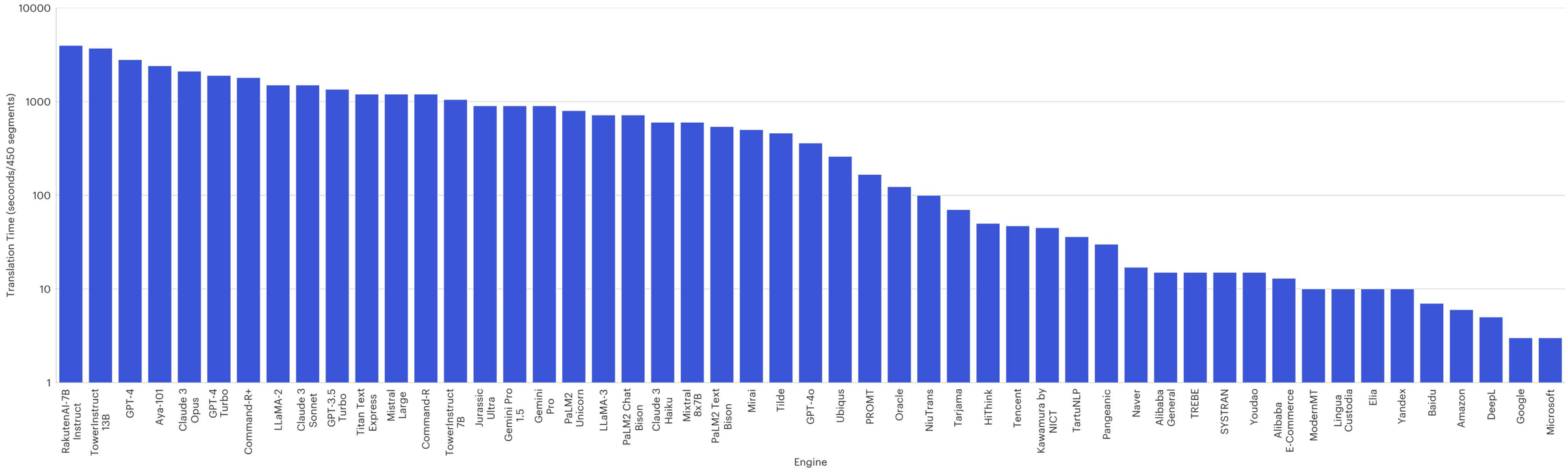


* Multilingual Large Language Models' language support for pairs other than the ones in the current MT Report cannot be confirmed

5.3 LLMs are 50-1000 Times Slower than Specialized MT Models

provider translation time (logarithmic scale)

There is a 50–1000 times difference in translation time between LLMs and MT engines*



* Several MT Engines' speed was affected by limited quotas (Tilde, Mirai, PROMT, NiuTrans, Tarjama, HiThink, Kawamura by NICT, TartuNLP, Pangeanic)

5.4 Changes in Providers' Features

→ Google has given General Access to their new translation engine which leverages Google LLMs to tailor translations, Adaptive Translation.

→ DeepL has expanded its language offerings by adding Arabic to its list of supported languages. This addition marks a significant milestone for the company, as Arabic becomes the first right-to-left language available on their platform.

→ Azure AI Custom Translator welcomes Neural Dictionary, an impressive extension to their dynamic dictionary and phrase dictionary features, and gives access to direct model customization.

→ OpenAI introduces their new multimodal flagship model, GPT-4o, which shows performance comparable to GPT-4 Turbo but is much more efficient, generating text 2x faster and being 50% cheaper.

→ Google debuts Gemini 1.5 Flash, upgrades Gemini 1.5 Pro, and introduces new developer features, such as video frame extraction and parallel function calling.

→ Anthropic presents a new generation of Claude models, Claude 3, including Haiku, Sonnet, and Opus in ascending order of capability.

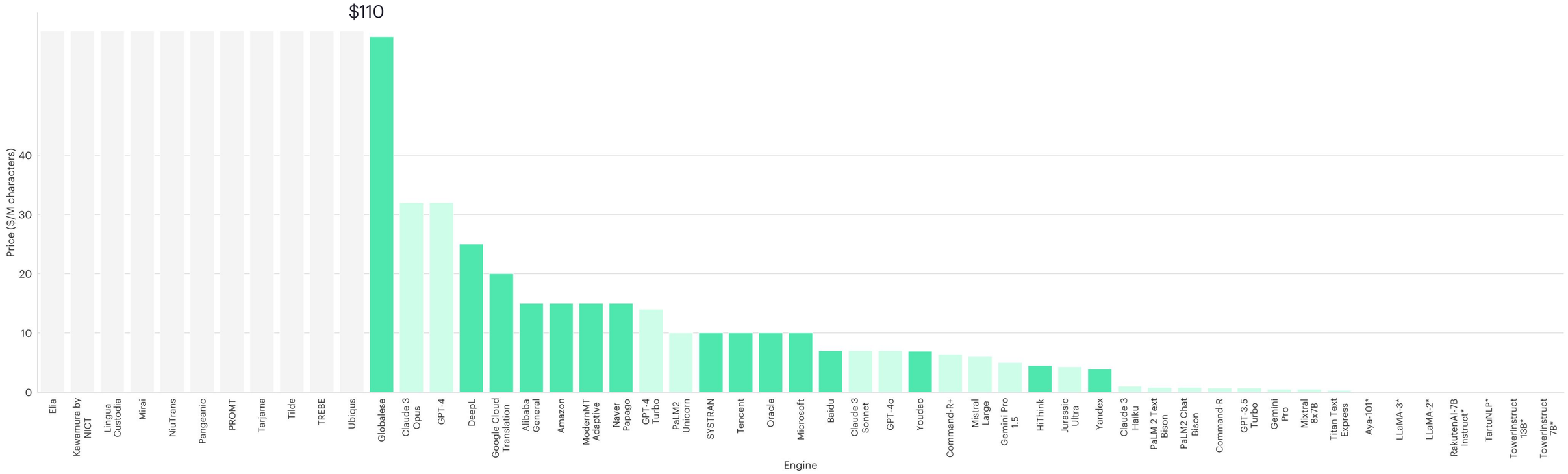
→ Yandex Cloud adds several new languages, among which there are three new to Intento: Yakut, Emoji, and Uzbek Cyrillic.

→ Microsoft expands to 20 Indian languages, with one of them, Bodo, being new to Intento.

5.5 LLMs Are Priced Lower Than MT Engines

- 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



* 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.

5.6 Public Pricing for Model Customization

	Amazon	Command R	Globalese	Google Vertex	Google v3	GPT-3.5 Turbo	LLaMA	Microsoft	Mistral AI	ModernMT Human in the Loop	Titan Text Express	SYSTRAN
Customization ($)	free	$8/M tokens	$50	$21.25/hour	$45/hour, $300 max	$8/M tokens	free	$10/M source + target characters (max. $300)	free	free	$0.008/ 1K tokens	by request
Hosting/month ($)	200 GB free parallel data storage. $0.023/GB per month for excess data	not specified	$5.50	not specified	free	free	free	$10	free	free	not specified	by request
Translation ($/M characters)	$60	$2.10	$110	from $0.10	$80	$3.20	free	$40	free	$50	$0.30	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.

5.7 Rapid Rise of Independent Cloud Vendors as Number of LLMs Grows 2.5 Times

Commercial 50

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, SAP, Sogou, Systran, Tarjama, Tencent, Tilde, Ubiqus, Unbabel, TREBE, XL8, Yandex, YarakuZen, Youdao

Preview / Limited 5

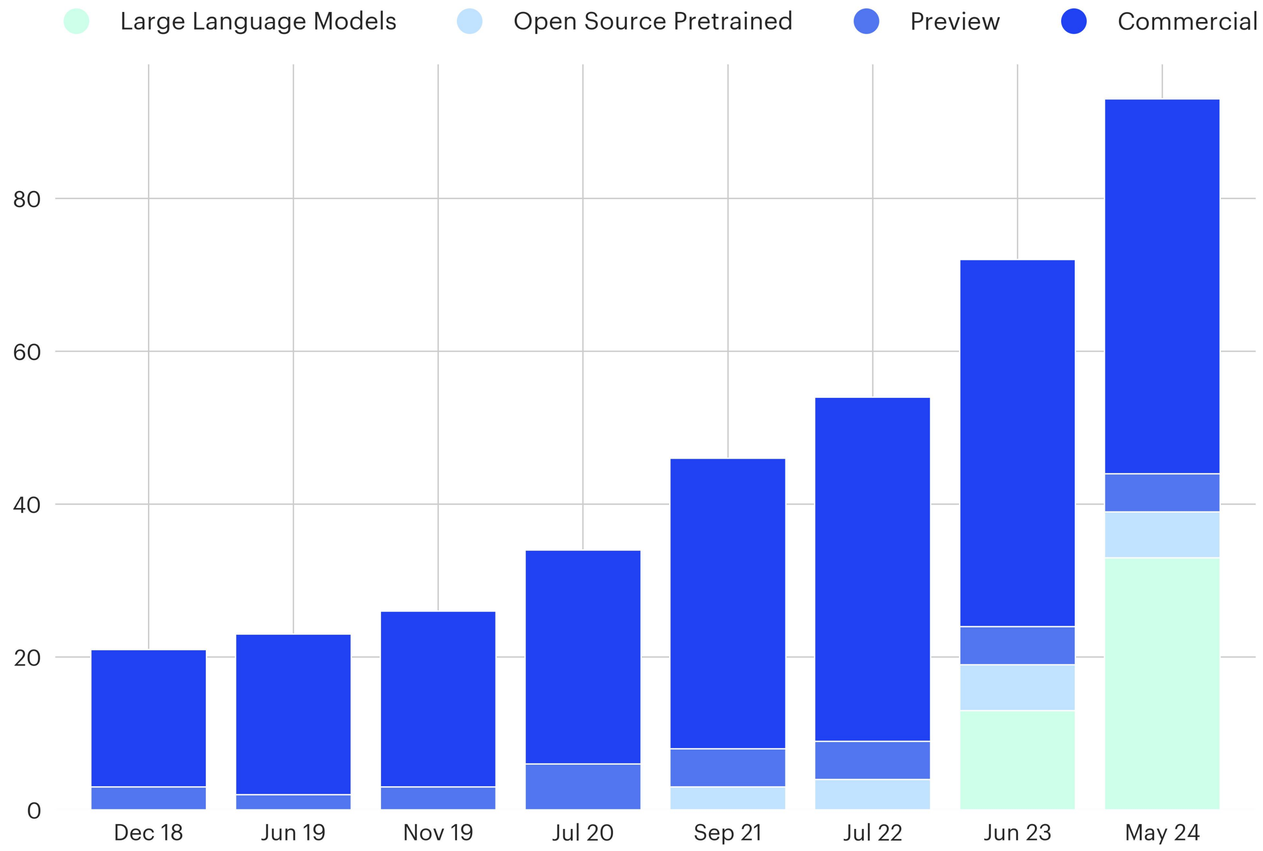
eBay, Kakao, QCRI, Tarjama, Birch.AI

Open Source Pretrained 6

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

Large Language Models 33

01.AI, AI21, Alibaba, AllenAI, Anthropic, BAAI, Baichuan, BAIR, BigScience, Cerebras, Cohere, DataBricks, Deci, EleutherAI, Google, HuggingFace, Lianjiia Tech, LLM360, LLMZoo, Meta AI, Microsoft, Mistral, MosaicML, OpenAI, Preferred Networks, Salesforce, SiloAI, Snowflake, Stability AI, Stanford, StatNLP, TII (UAE), X.ai



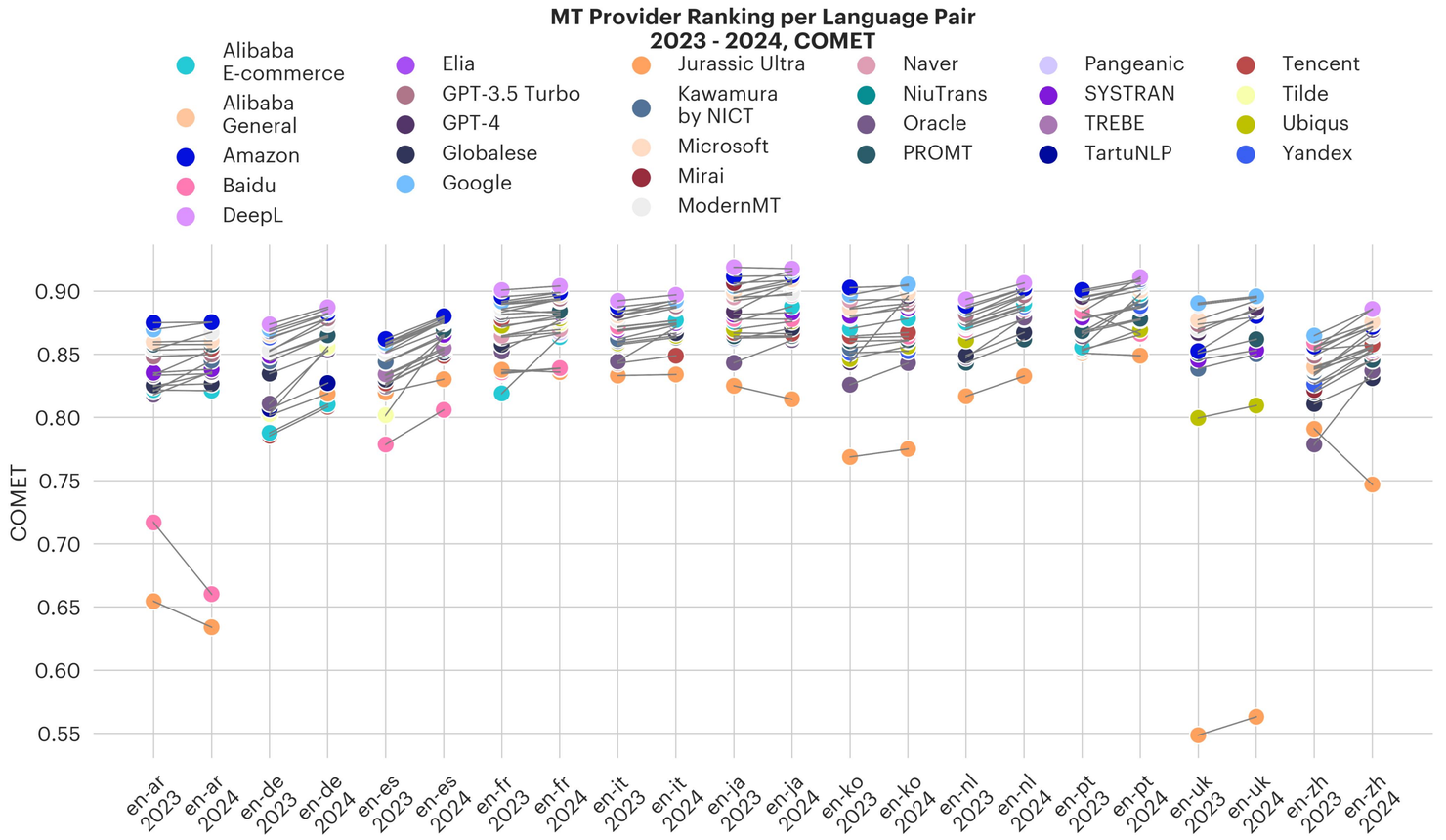
The new engines are highlighted in blue

5.8 Most Models Improved COMET-wise compared to 2023

→ Most providers have significantly improved score-wise.

→ Baidu has lower COMET scores than in the previous year in several pairs. In en-ar, which has the biggest score drop, we observe some moderate to severe mistranslations.

→ Jurassic Ultra, although having drastically improved in several pairs, shows decreased quality in en-ar, en-zh, en-ja, most likely to significant model updates that have happened over the year.



* For certain graphs, COMET was used in place of Intento LQA because the LQA analysis was limited to the top-performing engines, while COMET allowed for evaluation across all engines

5.9 Open Source Pre-Trained Models

Neurotõlge by TartuNLP

Neurotõlge is a multidirectional machine translation engine developed by the NLP lab at the University of Tartu. Among several high-resource languages, the engine supports several low-resource languages from the Finno-Ugric language family.

License: MIT License

Aya-101, Command R by Cohere

The Aya model is a massively multilingual open-source generative language model that follows instructions in 101 languages of which over 50% are considered as lower-resourced.

License: Apache-2.0

Command R is a 35B parameter generative model optimized for reasoning, summarization, question answering, and multilingual generation in 10 languages.

License: CC-BY-NC

Mixtral 8x7B by Mistral AI

Mixtral 8x7B is a generative sparse mixture of experts model (SMoE) with open weights, achieving high quality performance in several languages.

License: Apache 2.0

Llama-2, Llama-3 by Meta AI

Llama family of models contains Llama-2 and Llama-3 open-source large language models of different sizes designed for research, development, and innovation in the AI field.

License: Llama2 and Llama3

TowerInstruct 13B, TowerInstruct 7B by Unbabel

Tower family of models includes open-weight multilingual LLMs of different sizes for translation-related tasks, ranging from pre-translation tasks to translation and evaluation tasks. Build on top of Llama-2, they come in two sizes, 7B and 13B.

License: CC-BY-NC-4.0

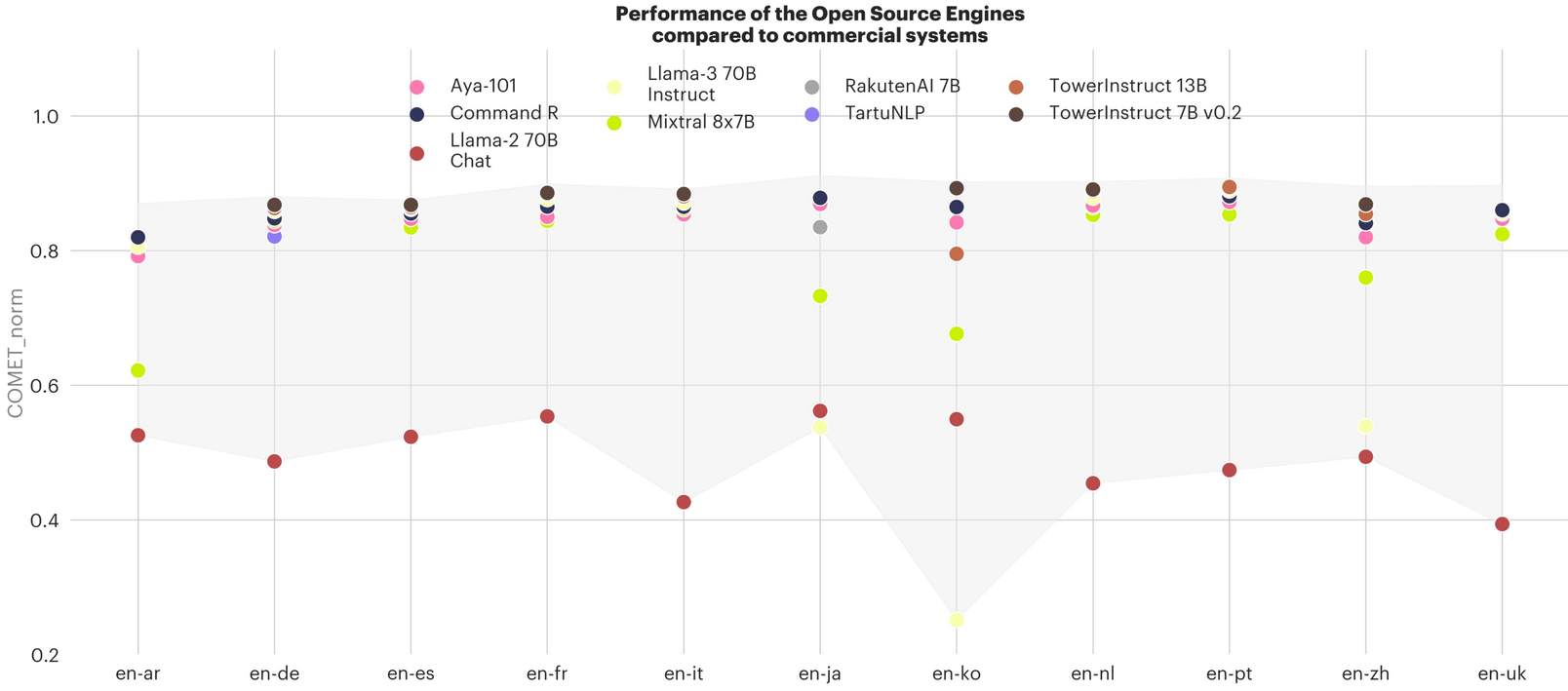
RakutenAI 7B by Rakuten

RakutenAI 7B Instruct build upon the Mistral model architecture and is based on Mistral 7B pre-trained checkpoint. It excels at Japanese language understanding while remaining competitive in English, outperforming similar models.

License: Apache-2.0

5.10 Several Open Source Models Deliver Impressive Results*

- → TowerInstruct 7B v0.2 and TowerInstruct 13B v0.1 outperform all other open-source models for most language pairs.
- → Translation in Arabic is the hardest for open-source models to tackle, with Command R scoring the highest but underperforming compared to commercial engines.
- → Llama-2 severely underperforms compared to the rest of the engines, while its successor Llama-3 achieves much higher scores in all language pairs except for en-ja, en-ko, and en-zh.



* For certain graphs, COMET was used in place of Intento LQA because the LQA analysis was limited to the top-performing engines, while COMET allowed for evaluation across all engines

5.11 Large Language Models

GPT by OpenAI

GPT-4o, GPT-4 Turbo, GPT-4 and GPT-3.5 Turbo are a diverse set of models with different capabilities and price points. All of them are fine-tuned for chat conversations, with GPT-4o and GPT-4 Turbo also having Vision capabilities.

Claude 3 by Anthropic

Claude 3 Haiku, Claude 3 Sonnet and Claude 3 Opus represent a series of LLMs with increasing levels of performance. The tiered approach allows users to choose the model that best suits their needs in terms of performance, speed, and cost-efficiency.

Titan Text Express by Amazon

Titan Text Express, exclusive to Amazon Bedrock, leverages Amazon's extensive AI and ML expertise. It is a high-performing text model supporting 100+ languages.

Gemini Pro by Google

Gemini is a family of generative AI models that are designed and trained to handle both text and images as input.

PaLM 2 by Google

Text Unicorn and Text Bison represent the PaLM 2 family of LLMs, offering enhanced capabilities in multilingual understanding, logical reasoning, and code generation.

Mistral Large by Mistral AI

Mistral Large, Mistral AI's flagship model, can be used for complex multilingual reasoning tasks, including text understanding, transformation, and code generation.

Jurassic Ultra by AI21

Jurassic Ultra, the flagship model of the Jurassic series, can tackle the most intricate language processing tasks and creating advanced generative text applications.

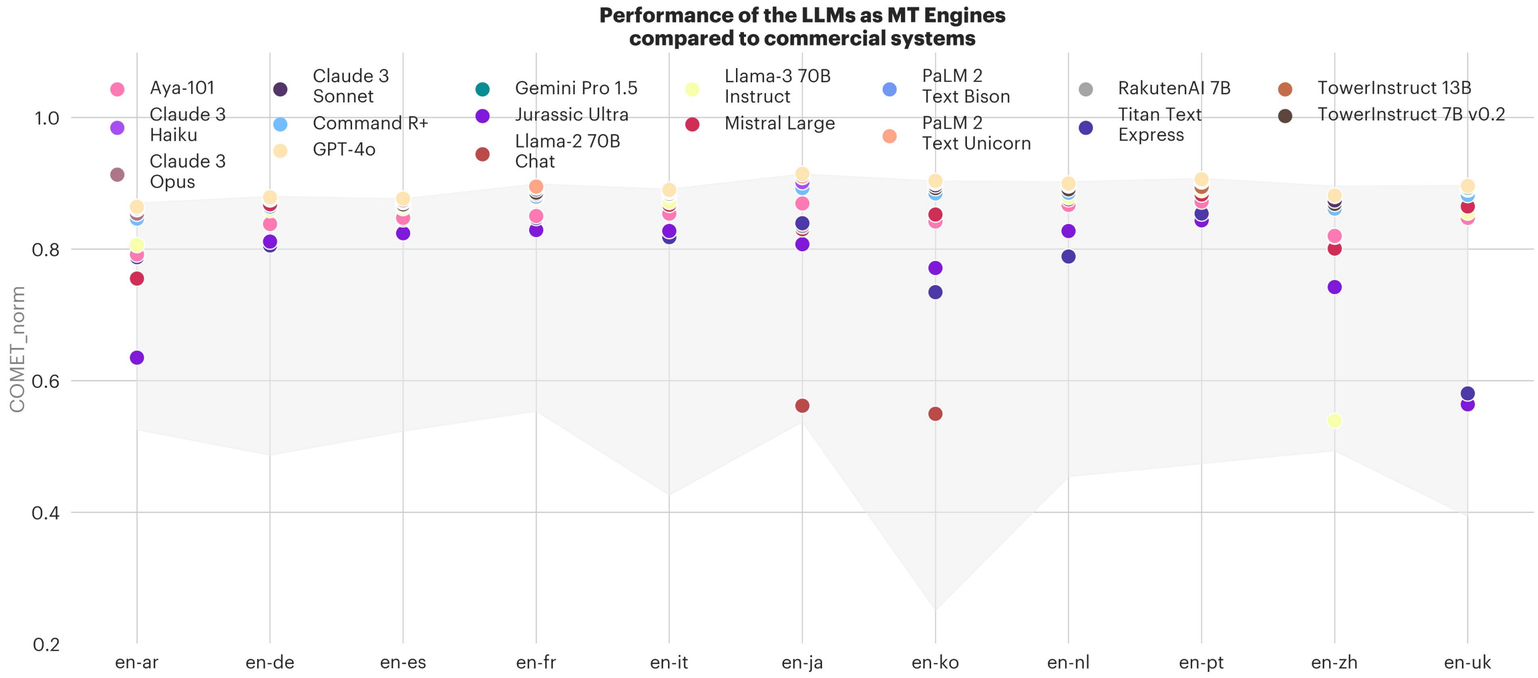
Other LLMs have been introduced earlier as open-sources models

5.12 Large Language Models Achieve Remarkable Scores in Top Tiers*

→ GPT-4o, which has recently been present as the new flagship model of OpenAI, consistently appears in the 1st tier and scores higher** than other OpenAI models.

→ PaLM2 Text Unicorn shows the highest results in en-fr out of all LLMs.

→ Models from the Llama family, Jurassic Ultra, and Titan Text Express generally achieved lower scores in comparison to the other models evaluated in this report.



* For certain graphs, COMET was used in place of Intento LQA because the LQA analysis was limited to the top-performing engines, while COMET allowed for evaluation across all engines

** Considering the number of Large Language Models, we have only chosen the best out of each model family by language pair to be present on the graph.

6. Takeaways

6.1 Key Conclusions (1/2)

1. Large Language Models are changing the Machine Translation landscape

Since the 2023 Report, we noticed two new vendors with pre-trained MT models. The growth in the number of new LLM providers has been more substantial. We currently track 94 vendors, 33% of them are LLM vendors (it was just 18% in 2023).

2. LLMs help with the linguistic quality analysis

To enhance our analysis, we have introduced Intento LQA, an LLM-based DQF-MQM-based metric, and combined it with the established COMET framework. It provides more insights into quality issues and enables a more accurate model comparison.

3. LLMs rapidly carve out their share of the market

This time, we assessed a total of 52 engines, out of which 24 were Large Language Models. It's not just about having more LLMs: we have more of them among best models, too. 55% of all top-performing models are LLMs (it was 17% in 2023). The largest LLM is not always the best for translation (even from the same provider), hence we see multiple models from the same family.

4. The quality landscape is quite complex

In 25% of all cases, LLMs are significantly better than any MT. More in Colloquial, Education and Entertainment. In 12% cases, MT is better than any LLM. More in English to Arabic, and in IT and Legal domains.

5. LLMs are much cheaper and much slower

LLMs (with simple prompts) are 10-100 times less expensive than MT systems and the price is not correlated with the quality. However, they are 50-1000 times slower, so we provide a separate rating for real-time systems. On average, the real-time requirement comes at 11% penalty in quality.

6. Few languages and domains are harder than others

Among the analyzed pairs and domains, the Colloquial domain and English-Arabic pair carry the most critical issues due to translation complexity. Overall, translations into Japanese and Korean have clear leaders for every domain except IT and Legal, emphasizing the importance of careful model selection.

6.1 Key Conclusions (2/2)

7. Number of supported languages didn't grow much

There are 191,010 unique language pairs across all MT systems. Among them, four new languages and 1,000+ new language pairs, but it's just a 0.5% growth since 2023. Although technically LLMs should support all languages, we see that the quality varies a lot, so it deserves a separate study.

8. 19 best-performing engines overall and 9 best in real-time scenario

Among the 9 domains and 11 language pairs analyzed, 18 MT engines and LLMs emerge as the top performers. When LLMs are excluded due to their high latency, 9 MT engines demonstrate the best performance across the board.

9. Open-source LLMs are generally in the 2nd tier

While the performance of open-source LLMs like TowerInstruct 7B v0.2 or Command R approaches top-tier commercial engines, the majority of open-source LLMs produce lower-quality translations due to their more limited multilingual capabilities compared to their commercial counterparts.

10. Few models are best at translation

Out of multiple tested LLMs, we see only models from Anthropic, Cohere, Google, and OpenAI in leaders. The number of leading MT vendors has also reduced in 2024. This could be attributed to a more detailed quality analysis.

11. MT and LLMs make similar translation errors

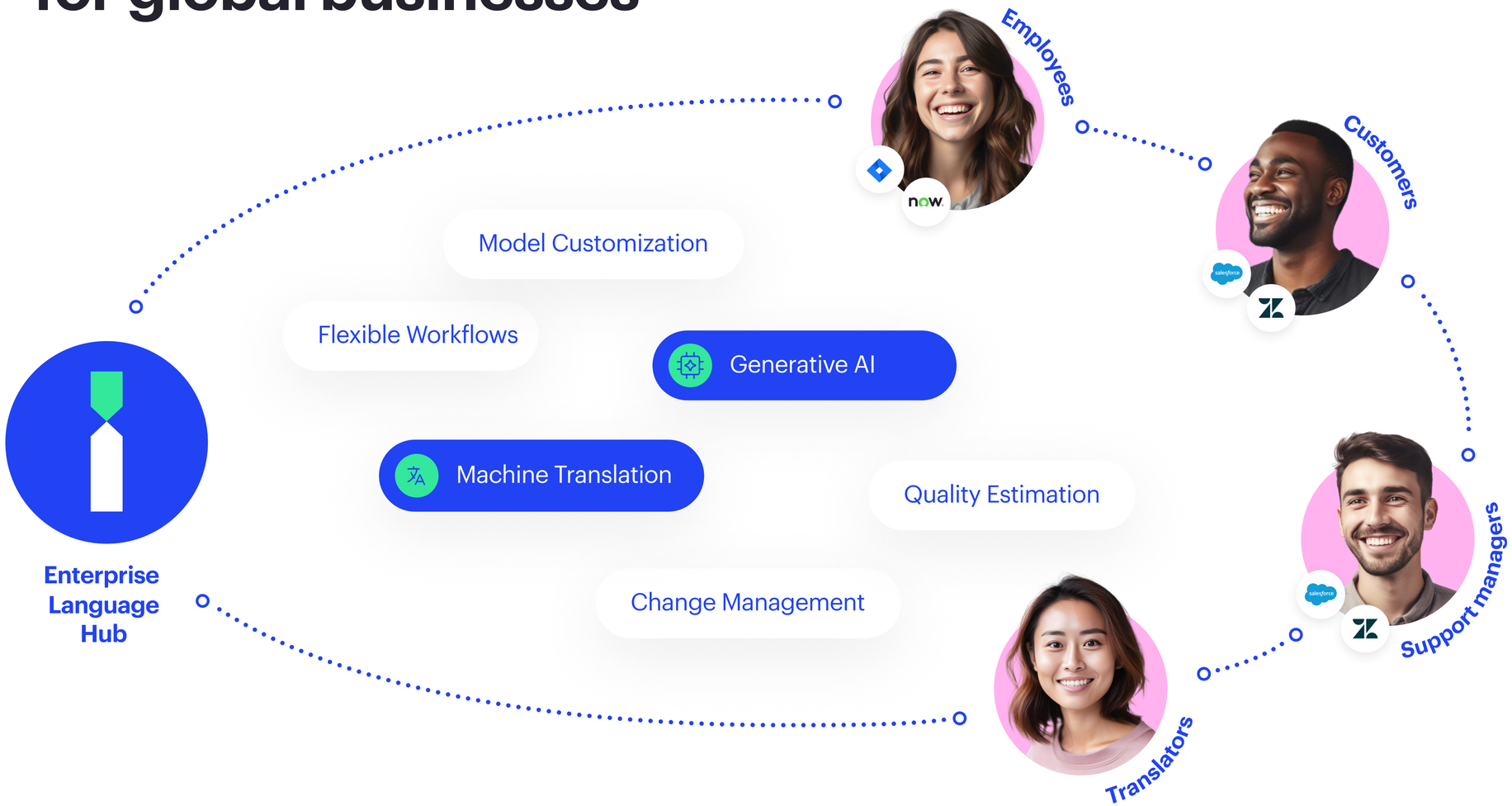
We did not see much difference in translation errors between MT engines and LLMs. However, some of the systems have certian perks. Some systems produce more grammatical errors than others, some can be more awkward in Asian languages; one of the models produces more partially untranslated segments, while yet another is prone to additions.

12. Customization improves quality above this report

Both MT and LLMs can be trained on translation memories and improved with glossaries. LLMs additionally can benefit from prompt engineering and RAG*. These tools can eliminate significant amount of the errors found in this report. Use them.

* Retrieval Augmented Generation

6.2 Intento – Machine Translation and multilingual Generative AI platform for global businesses



Trusted by the global enterprise

PROCORE

Government of Canada

vmware

Agilent

playrix

6.3 Unlock machine translation and generative AI across your entire company with an all-in-one, scalable platform

The right way to jumpstart your AI program

We customize and evaluate AI models for you, configure workflows and integrations, and provide ongoing maintenance.

Keep existing workflows running smoothly

Enterprise Language Hub integrates with the most popular software systems, so you can keep you existing human workflows in localization, document management and customer support.

Save up to 20x on translation and content production

Combine MT with automatic language skills, such as source content improvement and automatic revision, to save up to 95% on what you spend today.

Central, future-proof enterprise AI deployment

Build and manage AI workflows centrally, tapping into 40+ MT and GenAI providers, ready for any new tech the future brings. ISO-27001 certified.

Book a demo

6.4 Get maximum results from machine translation

Modern machine translation (MT) is powerful but struggles with imperfect source text and cannot leverage context.

Enterprise Language Hub overcomes these limitations by making source text more translatable before MT and adding context afterward. It can also use GenAI to automate human post-editing, saving up to 95% of translation costs.

Book a demo

intento

Source quality improvement

Change incorrect formatting, slang, and language errors before translation.

Up to

25%

less editing

Machine translation

We help pick the right MT model for each of your languages and tailor it to your terminology and glossary.

Up to

70%

less effort

Automatic post-editing

Apply your tone of voice, terminology, or other customized language features with generative AI.

Another

60%

less editing needed

6.5 MT Evaluation & MT Maintenance for hassle-free Enterprise MT

MT Evaluation

- → Data cleaning
- → Model training
- → Test sample translations
- → Model training analysis
- → LQA (sample review)
- → Final analysis

Learn how to build or improve your MT program

MT Maintenance

- → MT Performance Monitoring & Hot-Swap
- → Glossary updates
- → Model updates
- → MT Quality Monitoring
- → Post-editing Effort Analysis
- → MT Evaluation

Learn how to evolve your MT program over time

Fast and Safe

Only 5-6 weeks to get a winning MT engine with estimations for effort saved in post-editing and quality in real-time cases, such as support chats

Trusted

We run 15–20 MT Evaluation projects per month for global companies across industries under strict Security, Quality, and Data Protection requirements. ISO 27001 and ISO 9001 certified.

The State of Machine Translation

An independent multi-domain evaluation of MT engines

Commercially available pre-trained MT models

2261 Market St, #4273
San Francisco, CA 94114

inten.to

3655 Nobel Drive, Suite 520
San Diego, CA 92122

e2f.com

Appendix A

A.1 Best scores per Domain (BLEU)

A.1 Best scores per Domain (BLEU)

→ In the past, we were often asked “OK, but what are the BLEU scores”? Today, it’s commonly accepted that one should not use BLEU score at all. However, since you’ve asked for it, we decided to give you the highest SacreBLEU scores in each combination of domain and language pair.

→ There’s no statistical significance test as BLEU is a corpus-based score.

→ Please keep in mind that BLEU, as a corpus-level score with a number of parameters, is not comparable not only across different languages, but also across different datasets and different BLEU implementations.