intento e2f

The State of Machine Translation 2022

An independent multi-domain evaluation of MT engines

31 MT Engines

11 Language pairs

9 Content Domains

Disclaimer

July 1—July 28, 2022

The MT systems used in this report were accessed from July 1 to July 28, 2022. Some of these systems may have changed since then.

Automatic scoring

This report demonstrates the performance of those systems exclusively on the datasets used for this report (see slide 12) using semantic similarity scores. The final MT decision requires Human LQA and depends on each specific use case.

Tailored Dataset

Data for all domains were collected in English from publicly available datasets and translated by e2f into 11 languages. The selected MT providers could not have had access to such data in the past for training their models.

Plain Text Only

The evaluation was done on plain text data. We often see different results for tagged text (like those found in CAT/TMS systems) for some MT vendors and language pairs due to imperfect inline tag support.

Valid for a Specific Dataset

Normally, we run multiple evaluations for our clients using various language pairs and domains, and observe different MT system rankings than those provided in this report.

There’s no “best” MT system

MT performance depends on how similar your data is to the data used to train the vendors’ models, as well as their algorithms.

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.

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

31 Machine Translation Engines evaluated

11 Language pairs

English to →

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).

16 Machine Translation engines show the best results for some language pairs and domains

Alibaba	Microsoft	Ubiqus
Amazon	ModernMT	XL8
Apptek	Naver	Yandex
Baidu	PROMT	Youdao
DeepL	SYSTRAN	
Google	Tencent	

Massive language expansion across all MT engines

125,075 unique language pairs

+26,000 compared to 2021 — and still growing

The machine translation market is growing. Since The State of MT 2021 report, 4 more vendors now offer pre-trained MT models, and several open-source pre-trained MT engines have become available.

We've evaluated 31 engines, including No Language Left Behind by Meta AI, which has just been made available to the public.

We chose COMET from among 6 metrics for a better correlation with human translation.

Many engines perform best for English to Spanish and Chinese. Legal, Financial, IT, and Healthcare require a careful choice of MT vendor, as few perform at the top level. Entertainment and Colloquial show relatively low scores, which may indicate the importance for customization there.

Engines from Meta AI perform in the 2nd tier of commercial systems, except for English to Spanish (1st tier), English to Chinese, and English to Japanese (low performance).

About Intento

Gartner
COOL VENDOR 2021

Translate 20x more on the same budget

Intento allows global enterprises to translate 20x more on the same budget. Its tools help evaluate, customize, and connect best-fit AI to existing software and vendors. With Intento, businesses can also monitor machine translation performance to continuously improve their entire machine translation program.

Trusted by Global Enterprise

SUBWAY · Government of Canada · Agilent · PROCORE · playrix · JOOM

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 six 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

Intento – Your MT Innovation Partner

MT Evaluation

Select the best-fit option among pre-trained models and custom models trained on your available data — optimal for your language pairs and domains.

2,909
Models evaluated by Intento

125k+
Language pairs available for evaluation

Evaluate best-fit MT for your data with Intento MT Studio or ask our experts for professional help.

MT Hub

Translate better, faster, and at scale. Keep your data secure and streamline your workflows.

70%
Less post-editing

97%
MT requires no human review

Equip your team with intelligent technology to create and translate content 4x faster, in real time.

MT Maintenance

Keep your engines at the forefront of cutting-edge technology. Stay up-to-date on new models and updates.

2,988
Crucial MT provider updates detected in 2021

347k+
Glossary terms added and checked for their impact on quality

Learn how to evolve your MT program over time.

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

31 Machine Translation Engines

11 Language Pairs

9 Content Domains

1. MT Engines

1.1 Machine Translation Landscape

1.2 Evaluated Machine Translation Engines

1.1 Machine Translation Landscape

Generic stock models

- AISA
- Alibaba
- Amazon
- AppTek
- Baidu
- DeepL
- eBay
- Elia
- Fujitsu
- Globalese
- Google
- GTCom
- IBM
- iFlyTek
- Kakao
- Kawamura powered by NICT
- Kingsoft
- Lesan
- Lindat
- LingvaNex
- Microsoft
- Mirai
- ModernMT
- Naver
- NiuTrans
- NTT
- Omniscien
- PangeaMT
- Process9
- Prompsit
- PROMT
- Rozetta
- RWS
- SAP
- Sogou
- SYSTRAN
- Tencent
- Tilde
- Ubiqus
- Vicomtech
- Yandex
- YarakuZen
- Youdao

Vertical Stock Models

- Alibaba
- Baidu
- Cloud Translation
- Microsoft
- NiuTrans
- Omniscien
- PROMT
- RoyalFlush
- SAP
- SYSTRAN
- Ubiqus
- XL8

Custom terminology support

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

Auto domain adaptation

- Amazon
- Globalese
- Google
- IBM
- KantanAI
- Microsoft
- ModernMT
- Omniscien
- Rozetta
- RWS
- SYSTRAN
- Yandex

Manual domain adaptation

- Alibaba
- AppTek
- Baidu
- Cloud Translation
- Omniscien
- PangeaMT
- Prompsit
- PROMT
- RWS
- SYSTRAN
- Tilde
- Ubiqus

1.1 Machine Translation Landscape

Generic Stock Models

Pre-trained models based on generic data without a specific domain, meaning that 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 provider, terminology can be used while training custom models or for adjusting machine translation results.

Manual Domain Adaptation

The user comes directly to a provider and requests a customized model in a particular domain.

Vertical Stock Models

Follows the same logic as Generic Stock Models, as users do not customize the MT models. However, they do fit under a specific domain, relying on the context surrounding a particular industry, such as Healthcare or Finance.

Auto Domain Adaptation

Provides an UI or and API to customize a pre-trained (baseline) model with data provided by users in an automated fashion.

1.2 Evaluated Machine Translation Engines

Customization options

- None
- TM
- Glossary
- Both

Engine	Product	Customization
AISA	Neural Machine Translation API	None
Alibaba	eCommerce MT	None
Alibaba Cloud	General	None
Amazon	Translate	Both
Apptek	Neural Machine Translation	None
Baidu	Translate API	Glossary
DeepL	API	Glossary
Elia	Elhuyarren itzult-zaile automatikoa	None
Globalese	Machine Translation	TM
Google Cloud	Advanced Translation	Both
GTCom	YeeCloud MT	None
IBM Watson	eCommerce MT	Both
Meta AI	NLLB x4	None
Microsoft	Language Translator	Both
ModernMT	Realtime	TM
Naver	Papago NMT Commercial	None
NiuTrans	Translation Cloud Platform	None
Pangeanic	Machine Translation API	None
PROMT	Cloud API	None
RoyalFlush	Finance Translation	None
Rozetta T-400	Machine Translation API	Both
SYSTRAN	PNMT	Both
Tilde	Machine Translation API	None
Tencent Cloud	TMT API	None
Ubiqus	Translation API	Both
Yandex	Translate API	TM
Youdao	Cloud Translation API	None
XL8	Machine Translation	None

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 segments suitable for research

Data samples to reproduce this study are available by request from e2f and Intento

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 Industry Sectors and Language Pairs

9 content domains per language pair

500 segments in 11 language pairs per domain

This year, we have identical segment coverage for all language pairs.

Available resources

	en–ar	en–zh	en–nl	en–fr	en–de	en–it	en–ja	en–ko	en–pt	en–es	en–uk
Colloquial	500	500	500	500	500	500	500	500	500	500	500
Education	500	500	500	500	500	500	500	500	500	500	500
Entertainment	500	500	500	500	500	500	500	500	500	500	500
Financial	500	500	500	500	500	500	500	500	500	500	500
General	500	500	500	500	500	500	500	500	500	500	500
Healthcare	500	500	500	500	500	500	500	500	500	500	500
Hospitality	500	500	500	500	500	500	500	500	500	500	500
IT	500	500	500	500	500	500	500	500	500	500	500
Legal	500	500	500	500	500	500	500	500	500	500	500

Color scale: 540, 520, 500, 480, 460

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."

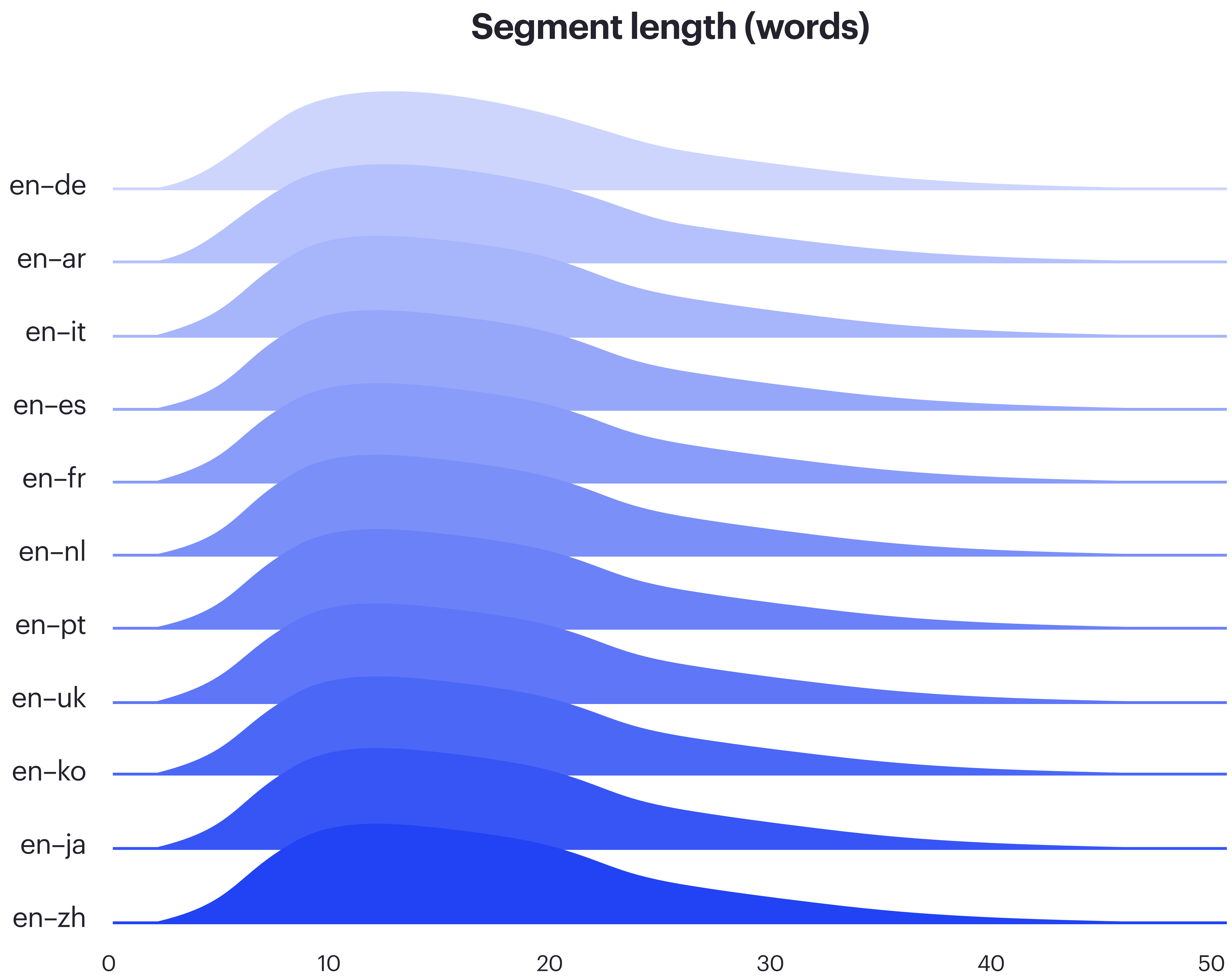
Colloquial

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

2.4 Sentence Length

→ The same segments were translated into 11 languages.

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



3. Evaluation Methodology

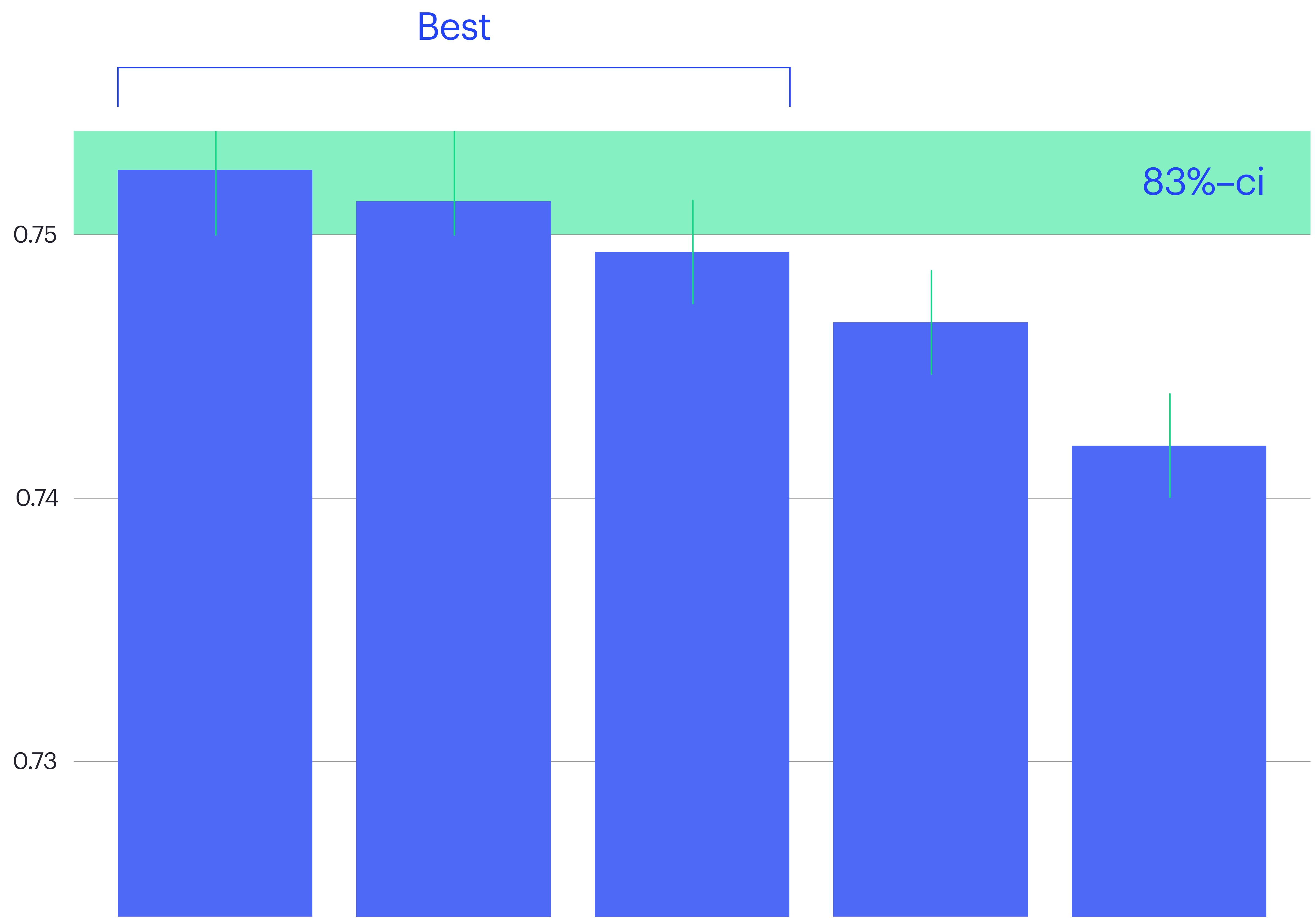
3.1 Evaluation Approach

3.2 What Scores to Use

3.1 Evaluation Approach

1. Rank MT engines based on a score showing distance from a reference human translation.
2. Identify a group of top-runners (BEST) within a confidence interval of the leader.

Using segment-level scores averaged across the corpus and an 83% confidence interval [1,2]



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.

3.2 What Scores to Use

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

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

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

PRISM

Semantic similarity

Evaluates machine translation as a paraphrase of a human reference translation. Penalizes both fluency and adequacy errors. Does not penalize paraphrases/synonyms. N/A for Korean.

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. See why we chose COMET as the main score.

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

4. Evaluation Results

4.1 Best MT Engines per Language Pair (COMET)

→ 6 MT engines are among the statistically significant leaders for 11 language pairs.

→ DeepL and Google cover the best options for all languages when domains are ignored.

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

→ Absolute values are not shown to avoid confusion, as the scores are not comparable across language pairs.

→ The domain and content type mix is differenent for every language pair (see the next slide) and largely influences this leaderboard.

Best MT engines by normalized COMET score*

en–ar	Google			
en–de	DeepL			
en–es	Amazon	DeepL	Google	Yandex
en–fr	DeepL			
en–it	DeepL			
en–ja	DeepL			
en–ko	Google	Naver		
en–nl	DeepL			
en–pt	Google			
en–uk	Google	Yandex		
en–zh	Google	Youdao		

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

4.2 Best MT Engines per Domain

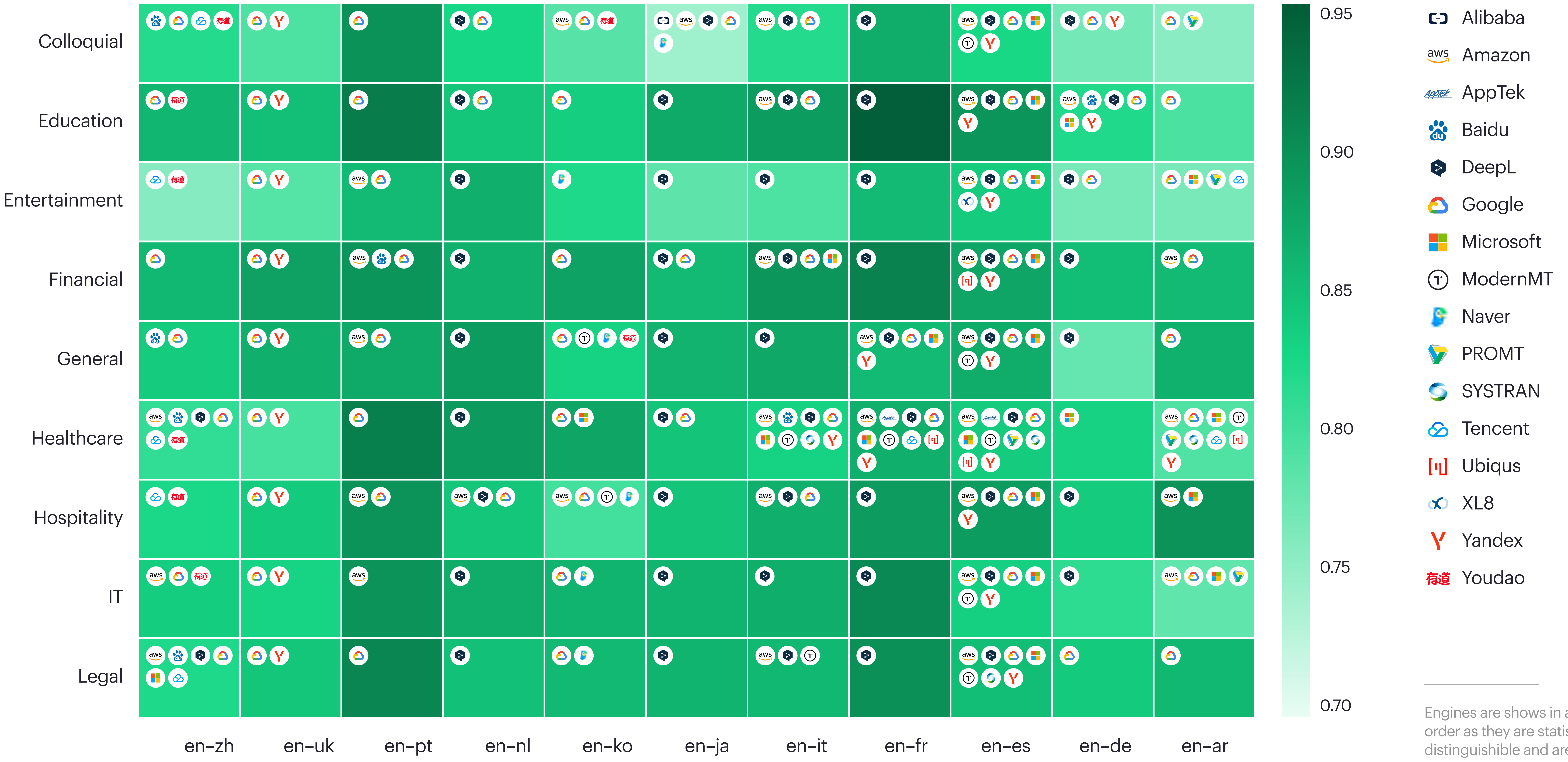
→ In the next slide, we show the best MT engines by normalized COMET score. Each square shows the best providers for a particular language pair in a specific domain. The color of the square shows the achievable MT quality for this domain compared to other domains in this language pair.

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

→ For each language pair, the score values were normalized to the [0,1] range, hence it's not comparable between different language pairs.

→ MT vendors 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 MT engines by domain per normalized COMET score



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

4.2 Best MT Engines per Domain

→ 16 MT engines are among the statistically significant leaders for 9 domains and 11 pairs.

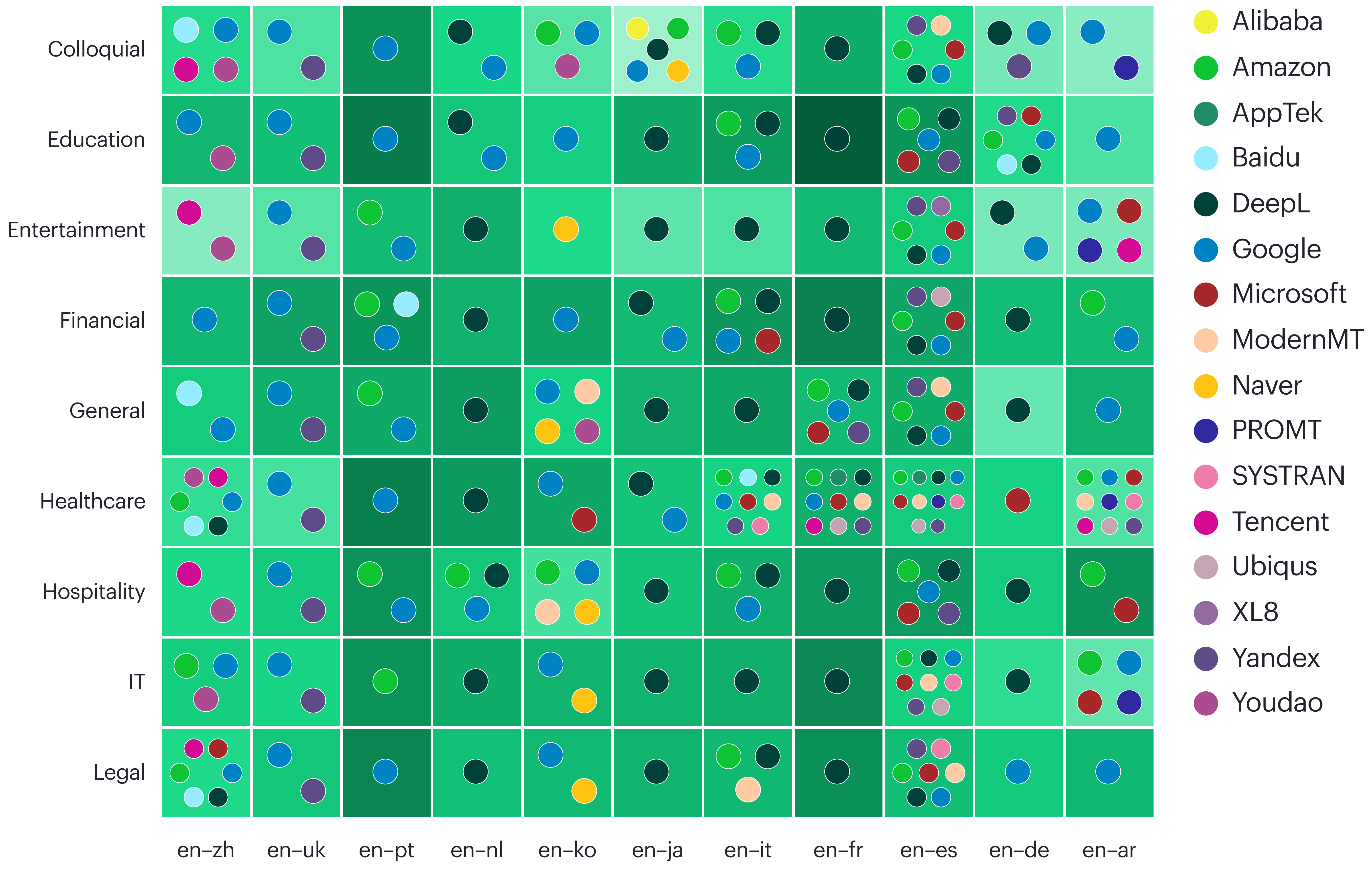
→ Many engines perform best with English to Spanish and Chinese.

→ Legal, Financial, IT, and Healthcare require a careful choice of MT vendor, as relatively few perform at the top level.

→ Despite having several comparable engines per language pair, Entertainment and Colloquial domains show relatively low scores, which may indicate the importance of customization.

→ In the Hospitality sector, COMET is higher than the BERTScore (see Slide 54), which may be due to how these models were trained; COMET was trained on post-edits while the BERTScore looks at the semantic similarities of texts.

Available quality and best MT engines by domain per normalized COMET score



4.3 Possible Minimal Coverage

6 MT engines provide minimal coverage* for all pairs and industries, 2–4 per domain.

Entertainment
DeepL, Google, Naver, Tencent

Hospitality
Amazon, DeepL, Google, Tencent

Healthcare
DeepL, Google, Microsoft

IT
Amazon, DeepL, Google

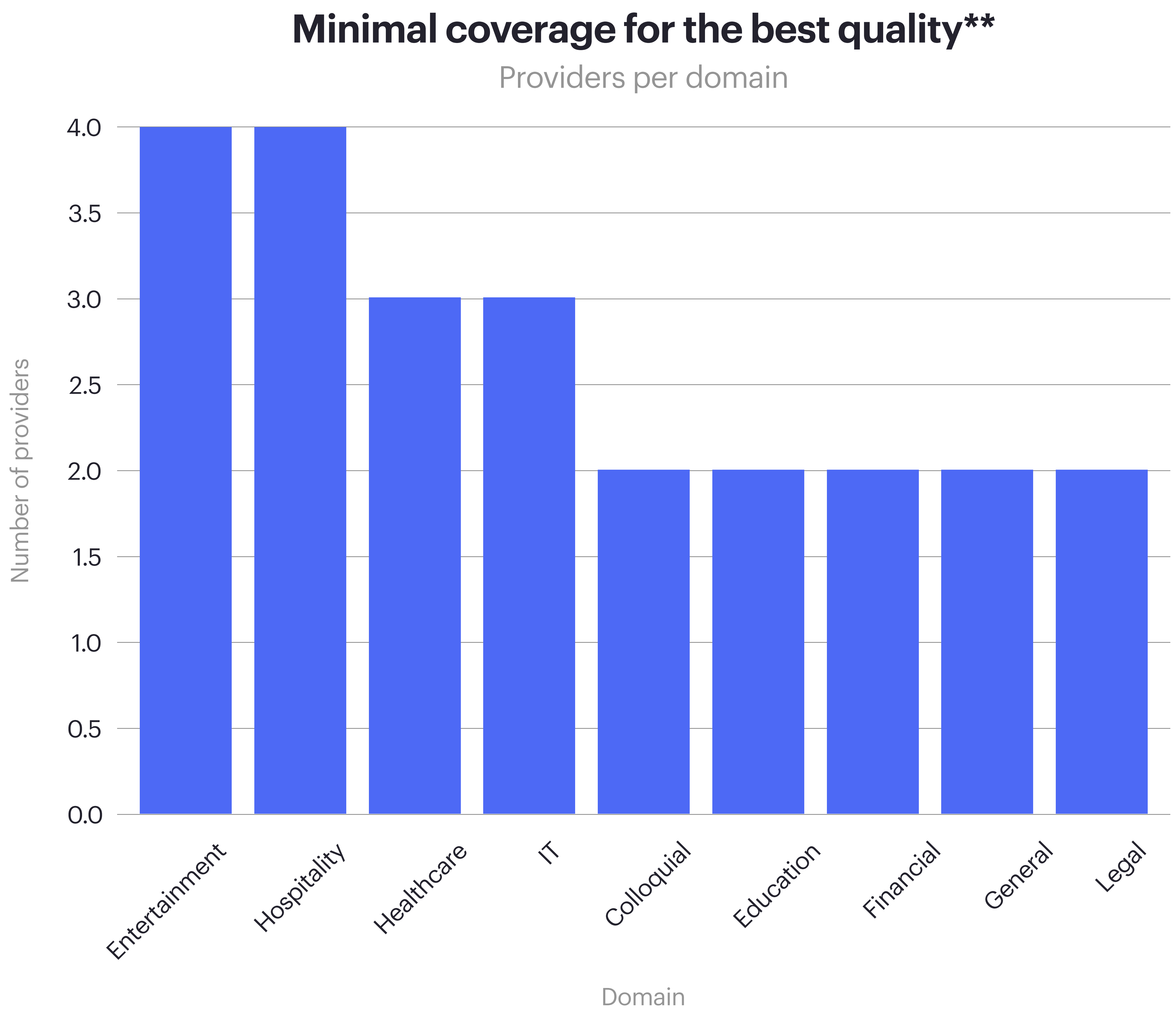
Colloquial
DeepL, Google

Education
DeepL, Google

Financial
DeepL, Google

General
DeepL, Google

Legal
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.

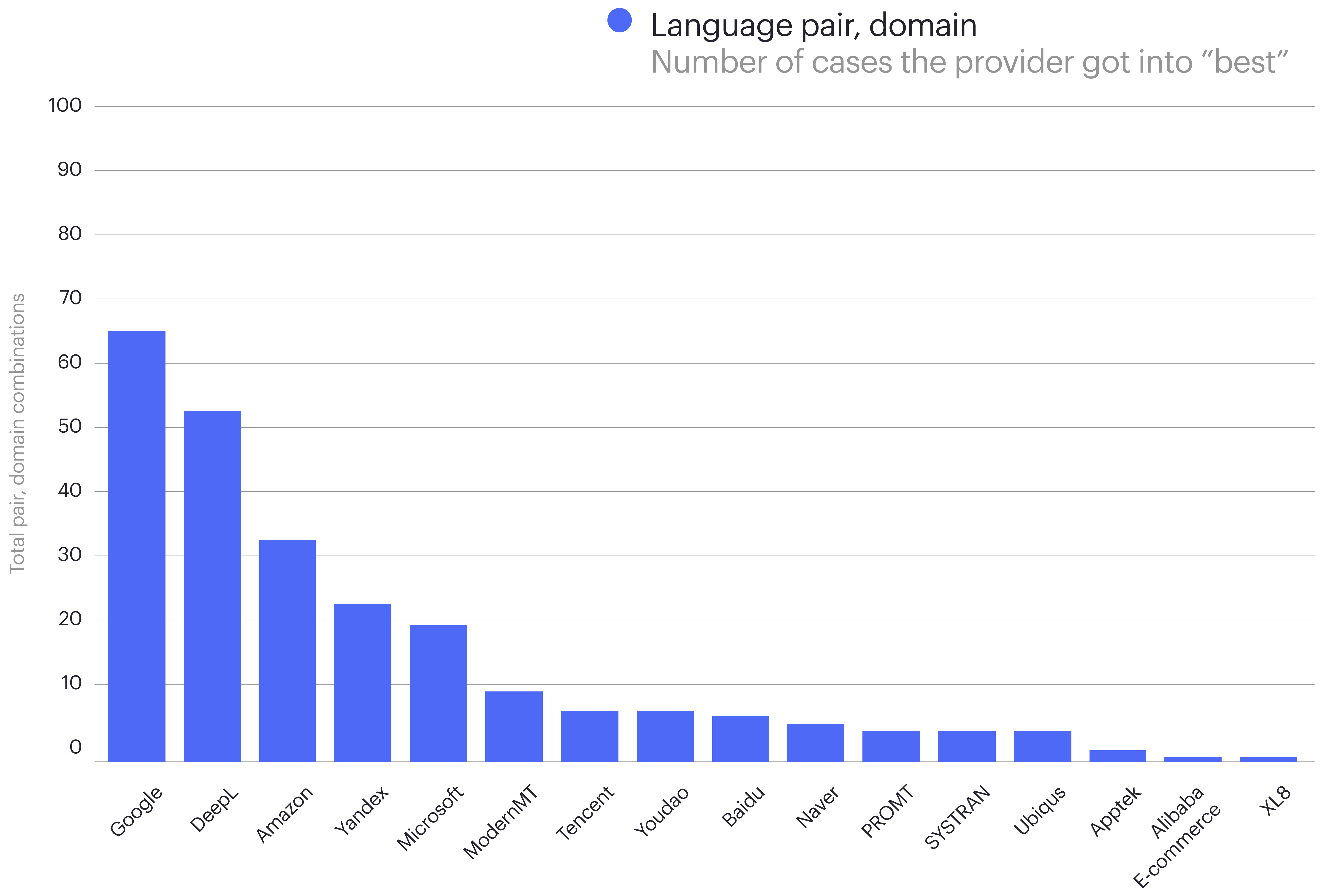
4.4 Top Performing MT Providers (COMET)

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

→ XL8 specializes in media localization; it was used in the Entertainment domain in en>es, en>fr, en>ko language pairs



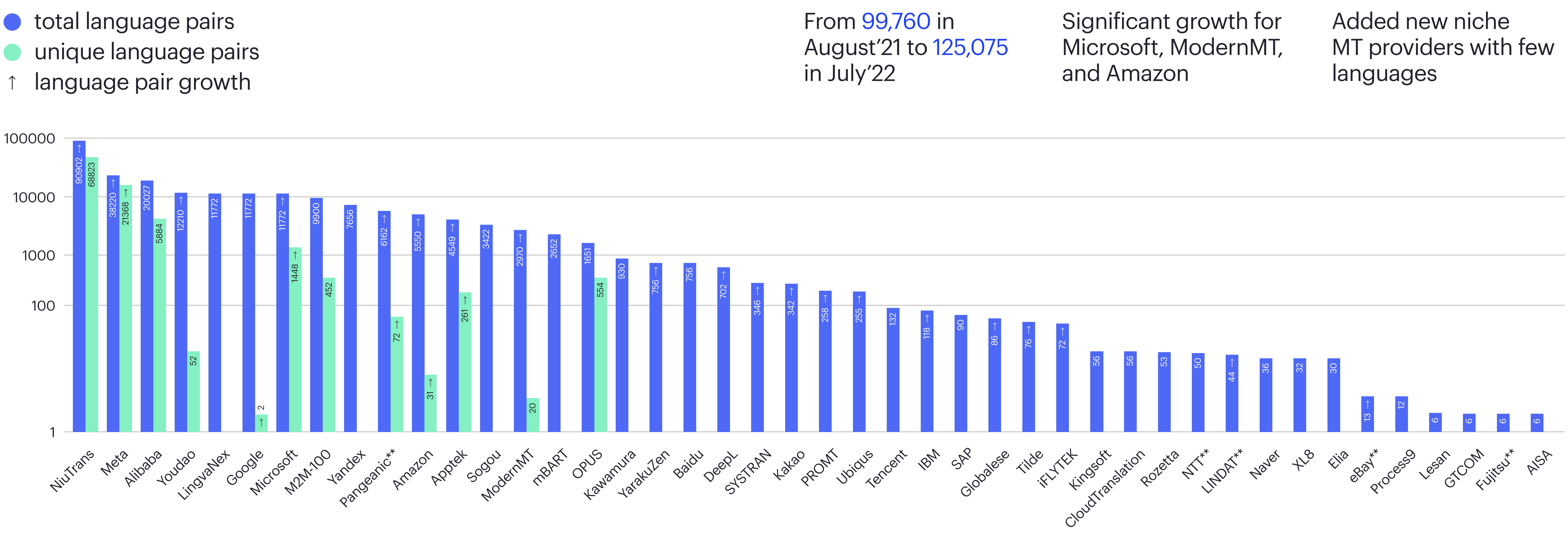
5. Miscellaneous

5.1 125,075 Language Pairs Across All MT Engines*

From 99,760 in August'21 to 125,075 in July'22

Significant growth for Microsoft, ModernMT, and Amazon

Added new niche MT providers with few 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).

5.2 Changes in Providers' Features

→ Amazon Translate added tone of voice support.

→ DeepL added two more languages to its glossaries feature:

- Italian <> English
- Polish <> English

They now have 14 language pairs that support glossaries.

→ At the end of 2021, Microsoft passed the 100 supported languages mark, bringing their overall language pair count up to more than 10,000. They are continuously adding languages, some of the last being Faroese, Somali and Zulu, and Basque and Galician.

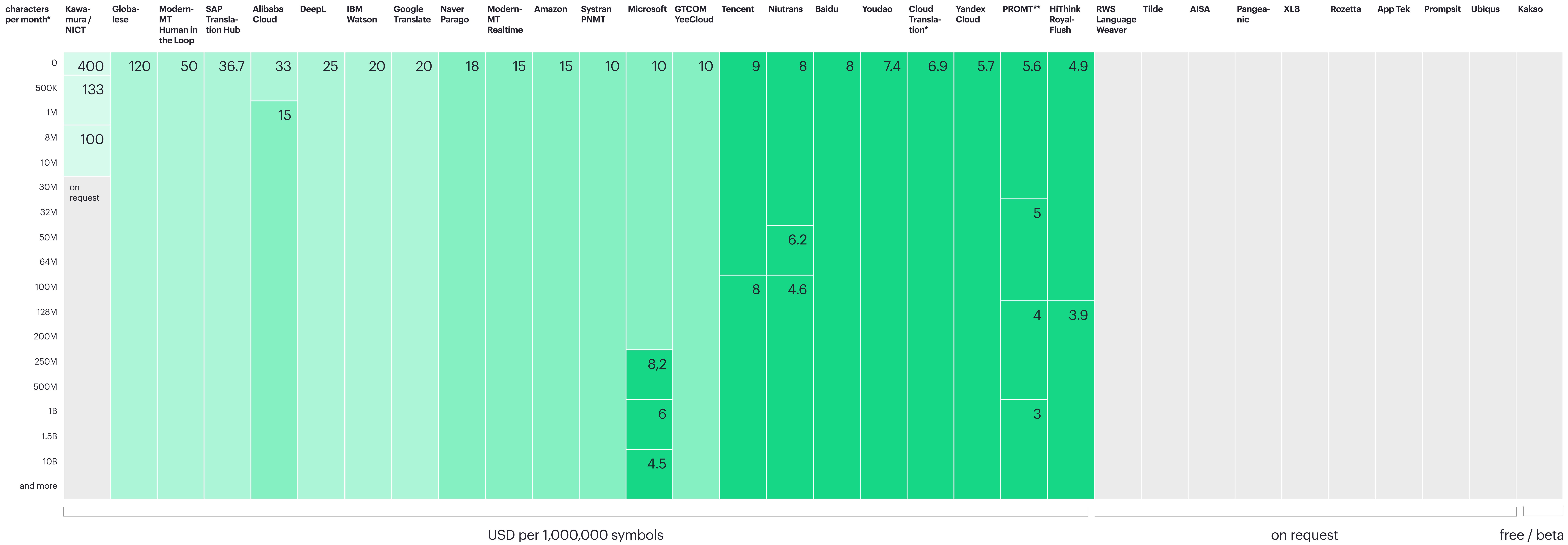
→ NiuTrans was added to Intento's list of providers, bringing the total number of language pairs up to 90,902, with 68,823 unique pairs.

→ A new large open-source model with 175B parameters called BLOOM has just been made avaliable for public use. BLOOM is able to generate text in 46 natural languages and 13 programming languages. For a lot of them, such as Spanish, French, and Arabic, BLOOM is the first language model with over 100B parameters ever created.

→ Meta AI has made public their No Language Left Behind models, which are stated to be particularly good for working with low-resource languages.

5.3 Public Pricing

USD per 1M characters***



* Volume estimation based on 4.79 characters per word.

** +20% for some language pairs.

*** Freemium volumes are not shown.

5.4 Independent Cloud MT Vendors with Stock Models

Commercial **45**

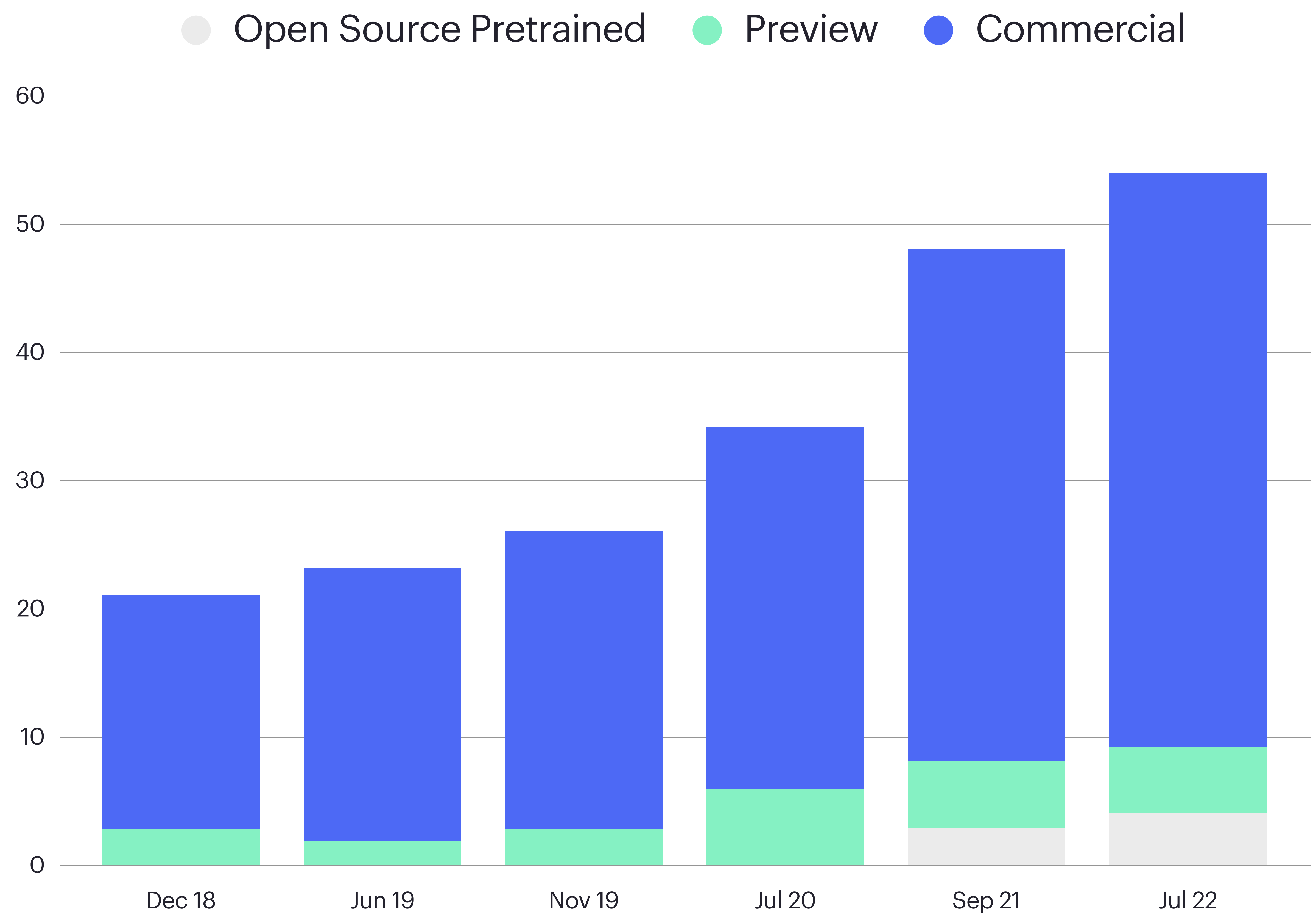
AISA, Alibaba, Amazon, Apptek, Baidu, CloudTranslation, DeepL, Elia, Fujitsu, Globalese, Google, GTCom, IBM, iFlyTek, RoyalFlush, Lesan, Lindat, Lingvanex, Kawamura / NICT, Kingsoft, Masakhane, Microsoft, Mirai, ModernMT, Naver, Niutrans, NTT, Omniscien, Pangeanic, Prompsit, PROMT, Process9, Rozetta, RWS, SAP, Sogou, Systran, Tencent, Tilde, Ubiqus, Vicomtech, XL8, Yandex, YarakuZen, Youdao

Preview / Limited **5**

eBay, Kakao, QCRI, Tarjama, Birch.AI

Open Source Pretrained **4**

NLLB by Meta AI, M2M-100, mBART, OPUS



The new engines are highlighted in blue.

5.5 Open Source Pre-Trained MT Engines

NLLB (Meta AI)

paper + code

No Language Left Behind (NLLB) is a project that open-sources models capable of delivering translations directly between a large amount of language pairs (200+ languages), including low-resource languages like Asturian, Luganda, Urdu, and others.

The creators open-source all evaluation benchmarks (FLORES-200, NLLB-MD, Toxicity-200), LID models and training code, LASER3 encoders, data mining code, MMT training and inference code, the final NLLB-200 models, and their smaller distilled versions.

The model is created by a large group of researchers at Meta AI, UC Berkeley, and Johns Hopkins University. In this report, we analyse 4 out of 5 publically avaliable models: 600M, 1.3B, 1.3B-distilled, and 3.3B.

OOS models evaluated in the 2021 MT Report

In the last year's "State of the Machine Translation", we evaluated three other open-source models: OPUS MT, M2M-100, and mBART50.

We have decided to omit them in this year's report as they have only shown results in the 2nd tier of commercial systems.

5.6 Open Source MT Performance (COMET)

→ NLLB by Meta AI mostly show performance in the 2nd tier of commercial systems.

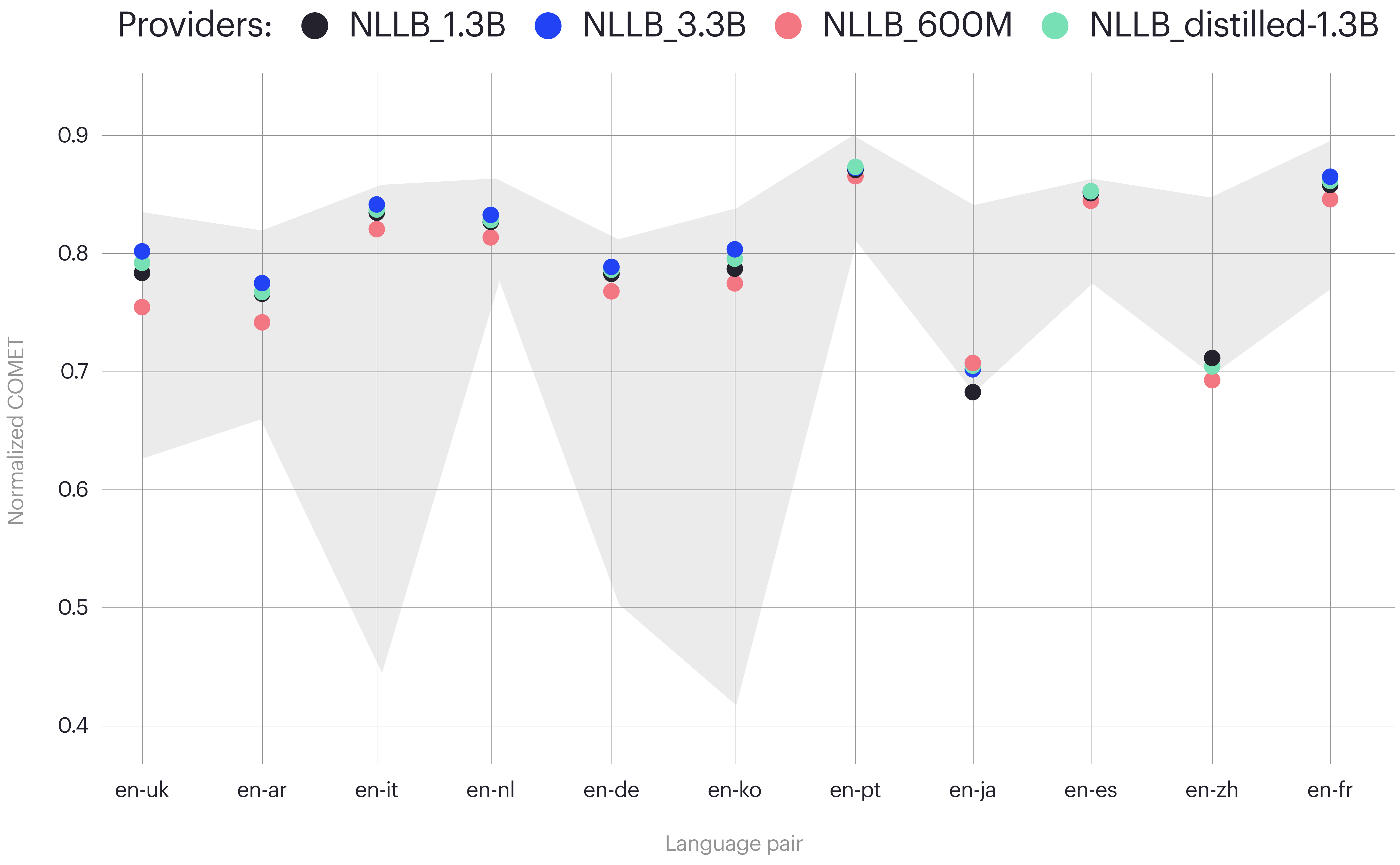
→ For en-es, NLLB scores are on par with the best commercial systems.

→ For en-zh and en-ja, the scores are quite low.

→ NLLB with 3.3B parameters leads for en-uk, en-ar, en-it, en-nl, en-de, en-ko, and en-fr.

→ NLLB with 1.3B parameters (distilled) leads for en-pt and en-es.

Performance of the Open Source Pretrained MT Engines compared to commercial systems



6. Takeaways

6.1 Key Conclusions

1. The MT market is growing: 45 vendors

Four more vendors offer pre-trained MT models since the 2021 MT Report, plus there are several open-source pre-trained MT engines available. We have evaluated 31 MT engines — among them is NLLB by Meta AI which has just been made public.

2. MT covers 125K language pairs

125,075 unique language pairs across all MT engines. 26K more than last year and still growing. The main contributors are Niutrans with their 90K language pairs, NLLB by Meta with 38K, and Alibaba with 20K.

3. 16 best performing MT Engines

16 MT engines are among the statistically significant leaders for 9 domains and 11 language pairs. 6 MT engines provide minimal coverage for all language pairs and domains, 2–4 per domain.

4. Open-source engines are in the 2nd tier

Open-source engines from Meta AI mostly perform in the 2nd tier of commercial systems, except for en-es (on par with top-tier systems) and en-zh & en-ja (much lower performance than commercial systems).

5. Four domains require a careful MT choice

Many engines perform best with English to Spanish and Chinese. Legal, Financial, IT, and Healthcare require a careful choice of MT vendor, as relatively few perform at the top level.

6. Two domains need more customization

Despite having several comparable MT engines per language pair, Entertainment and Colloquial show relatively low scores, which may indicate the importance of customization in these domains.

6.2 Intento — Your Compass in a Maze of Machine Translation

The MT market is constantly accelerating — and models need to be continuously re-evaluated to optimize localization budgets while ensuring the best translation quality.

Evaluate and customize MT with your dataset on many platforms at once with Intento MT Studio or ask our experts for professional help.

Book a demo

2,909

Evaluated models by Intento

125k+

Language pairs available for evaluation

6.3 MT Evaluation & MT Maintenance for a World-Class MT Program

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
→ Localization Checkup
→ 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.

Information Security Management System ISO 27001 Certified

CERTIFIED ISO 9001:2015 COMPANY

6.4 MT Hub. The Fastest Way to Translate 20x More

Select the best-fit machine translation for all your business needs with just a single contract.

Book a demo

Localization

Make your translators 70% more productive and translate more content, faster, on the same budget. Works in XTM, memoQ, and 15+ TMS.

Software Development

Help your international dev teams code and collaborate seamlessly no matter what languages they speak.

Customer Service

Achieve 24/7 real-time multilingual support with 97% user satisfaction in Salesforce Service Cloud, ServiceNow Service Portal, Helpshift, and Zendesk.

Community Content

Make all your community content readable and searchable in native languages. Verint Community, ServiceNow Community & Case and Knowledge Management.

Office Productivity

Make your digital workplace accessible for all employees and boost their global productivity.

6.4 MT Hub. Connect Best-Fit MT with Your Existing Software and Vendors

- TMS / CAT tools
- Knowledge Bases
- Community Portals
- Live Chats
- Browsers & Microsoft Office

+20

Book a demo

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 Choosing the Score

PRISM — unstable behaviour

We cannot use PRISM for the purposes of this report as we observe unstable behavior, with translations similar to the reference getting scores lower than some of the imperfect paraphrases, making comparing the mean scores problematic for high-performing engines. Also, it does not penalize non-translations and is not available for Korean.

BERTScore — commonly used

We also provide results for BERTScore, as it is one of the most commonly used machine translation quality metrics.

COMET — better correlation with human translation

A choice has to be made between BERTScore allowing omissive paraphrasing, and COMET penalizing context-dependent alternative translations. We have decided to go with COMET for this report, as it has a better correlation with human ratings and judgement.

Highest BLEU scores

We have also added a matrix with the highest SacreBLEU scores in Appendix D, as BLEU was the baseline for machine translation evaluation for decades.

See the comparison of hLEPOR, BERTScore, PRISM and COMET in Appendix A.

A.1 Choosing the Score

We've run a separate study on 15 language pairs and 21 unique MT models where we compared several metrics with human reviewers' judgement.

We found that in 10 out of 15 language pairs COMET has a better correlation with human ratings than other metrics, in 3 out of 15 language pairs BERTScore shows slightly better correlation, and in 2 language pairs based only on the data we currently posses both BERTScore and COMET show lower correlation results.

Please note that we have analyzed the post-editing case, and for other use cases, such as gisting or understanding MT, BERTScore may be better.

Pearson correlation in en-de

	rating	BERTScore	hLEPOR	TER	COMET
rating	1.00000	0.0423	0.0769	-0.0940	0.1585
BERTScore	0.0423	1.00000	0.7998	-0.7926	0.5894
hLEPOR	0.0769	0.7998	1.00000	-0.8921	0.4962
TER	-0.0940	-0.7926	-0.8921	1.00000	-0.5069
COMET	0.1585	0.5894	0.4962	-0.5069	1.00000

Pearson correlation in en-pt

	rating	BERTScore	hLEPOR	TER	COMET
rating	1.00000	0.0976	0.0684	-0.1191	0.1667
BERTScore	0.0976	1.00000	0.7840	-0.7709	0.5049
hLEPOR	0.0684	0.7840	1.00000	-0.9062	0.4256
TER	-0.1191	-0.7709	-0.9062	1.00000	-0.4276
COMET	0.1667	0.5049	0.4256	-0.4276	1.00000

Pearson correlation in en-nl

	rating	BERTScore	hLEPOR	TER	COMET
rating	1.00000	0.1482	0.1648	-0.1653	0.2881
BERTScore	0.1482	1.00000	0.8406	-0.8355	0.6019
hLEPOR	0.1648	0.8406	1.00000	-0.8876	0.4732
TER	-0.1653	-0.8355	-0.8876	1.00000	-0.5088
COMET	0.2881	0.6019	0.4732	-0.5088	1.00000

Pearson correlation in en-fr

	rating	BERTScore	hLEPOR	TER	COMET
rating	1.00000	0.1545	0.1463	-0.1838	0.2477
BERTScore	0.1545	1.00000	0.7897	-0.8421	0.6427
hLEPOR	0.1463	0.7897	1.00000	-0.8978	0.5995
TER	-0.1838	-0.8421	-0.8978	1.00000	-0.6158
COMET	0.2477	0.6427	0.5995	-0.6158	1.00000

Pearson correlation in en-es

	rating	BERTScore	hLEPOR	TER	COMET
rating	1.00000	0.0233	0.0202	-0.0258	0.1793
BERTScore	0.0233	1.00000	0.8233	-0.8315	0.4637
hLEPOR	0.0202	0.8233	1.00000	-0.9184	0.4570
TER	-0.0258	-0.8315	-0.9184	1.00000	-0.4499
COMET	0.1793	0.4637	0.4570	-0.4499	1.00000

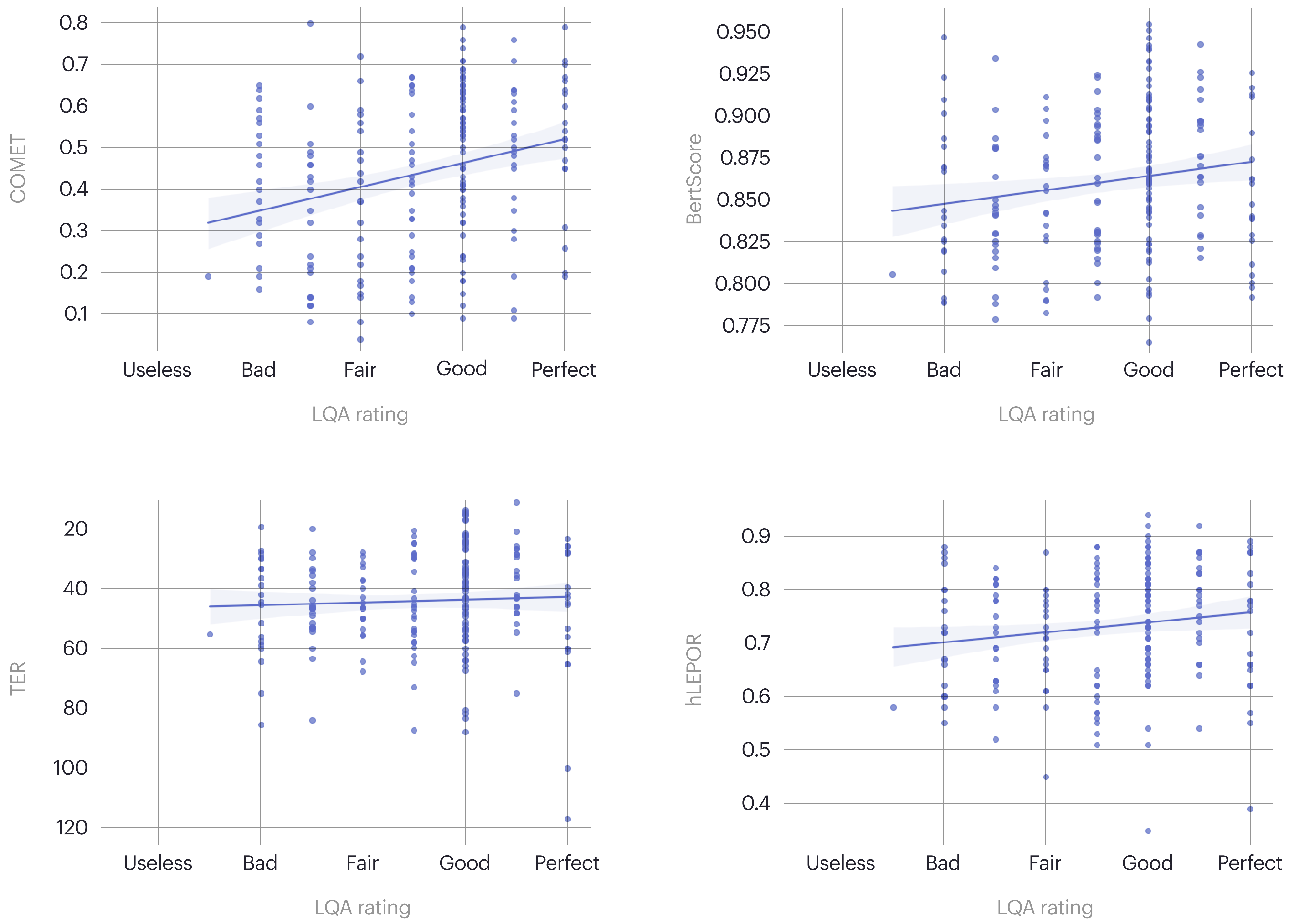
Pearson correlation in en-ko

	rating	BERTScore	hLEPOR	TER	COMET
rating	1.00000	0.1742	0.1537	-0.0489	0.2721
BERTScore	0.1742	1.00000	0.8068	-0.8200	0.4488
hLEPOR	0.1537	0.8068	1.00000	-0.7890	0.4676
TER	-0.0489	-0.8200	-0.7890	1.00000	-0.4098
COMET	0.2721	0.4488	0.4676	-0.4098	1.00000

See the comparison of hLEPOR, BERTScore, PRISM and COMET in Appendix B.

A.1 Choosing the Score

We have checked the correlations between several metrics and human judgment ratings. COMET has the best correalation in most cases.



See the comparison of hLEPOR, BERTScore, PRISM and COMET in Appendix A.

A.2 Going Forward with COMET

Our version of COMET is available for Intento customers via Intento API and MT Studio UI for Intento customers.

In the making of this report, wmt20-comet-da model in the COMET 1.0.1 package was used.

Reach us to learn more

→ Uses source segment, reference, and machine translation to find the machine-translated segment's correlation with human judgement.

→ Source texts and human translations often have different formatting, so we lowercase everything before applying COMET.

→ For every language pair, we have normalized COMET to fit [0,1] interval.

→ Does not reflect absolute quality level. Not comparable across language pairs.

→ We are grateful to Unbabel for releasing the COMET metric and appreciate Unbabel's support and guidance in configuring it.

See the comparison of hLEPOR, BERTScore, PRISM and COMET in Appendix A.

See the analysis for BERTScore in Appendix B.

Appendix B

B.1 Comparing hLEPOR and BERTScore

low hLEPOR + high BERTScore

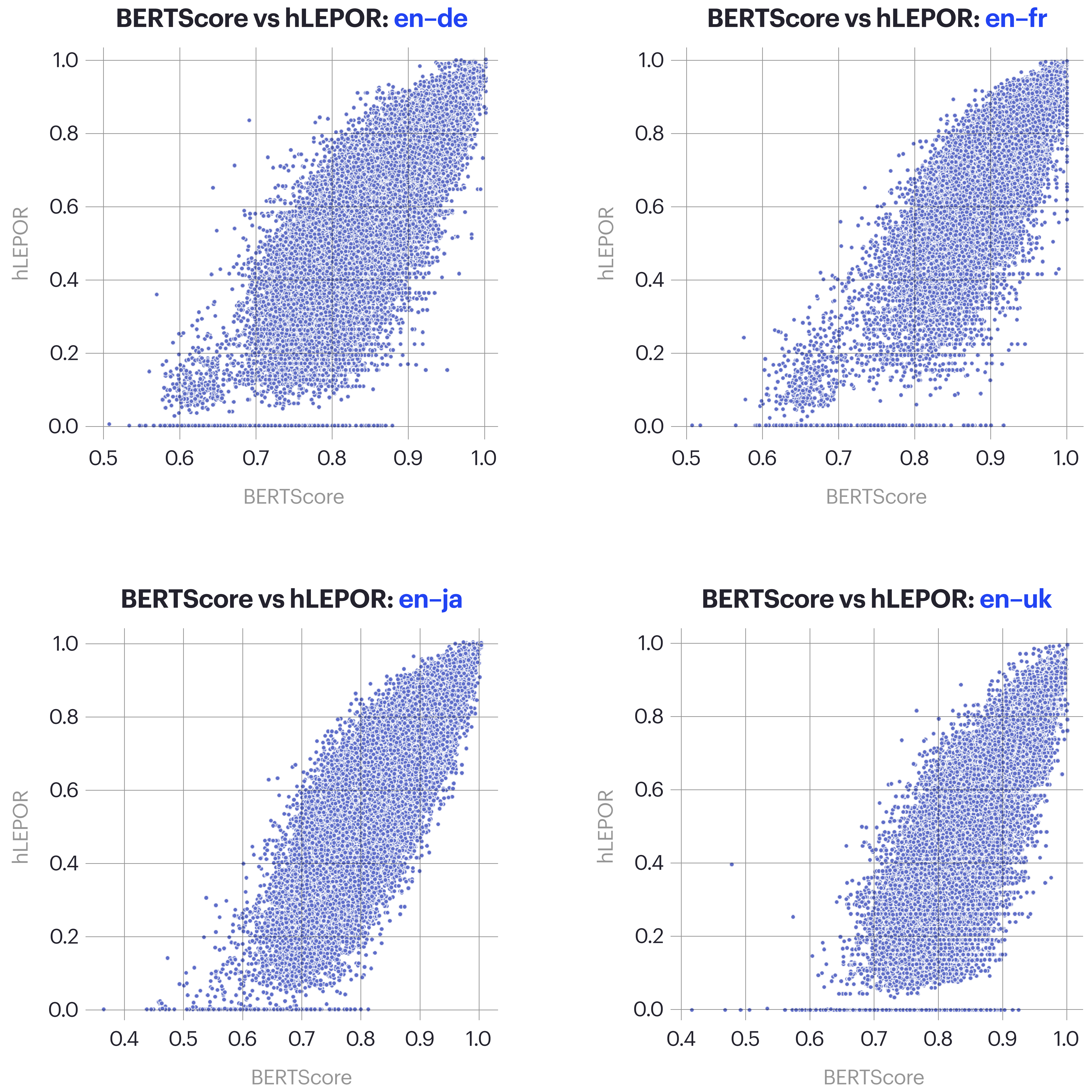
→ paraphrases / synonyms

→ minor differences in plurality between reference and MT

high hLEPOR + low BERTScore

→ mostly doesn’t exist

→ punctuation and spacing issues



B.2 Comparing hLEPOR and COMET

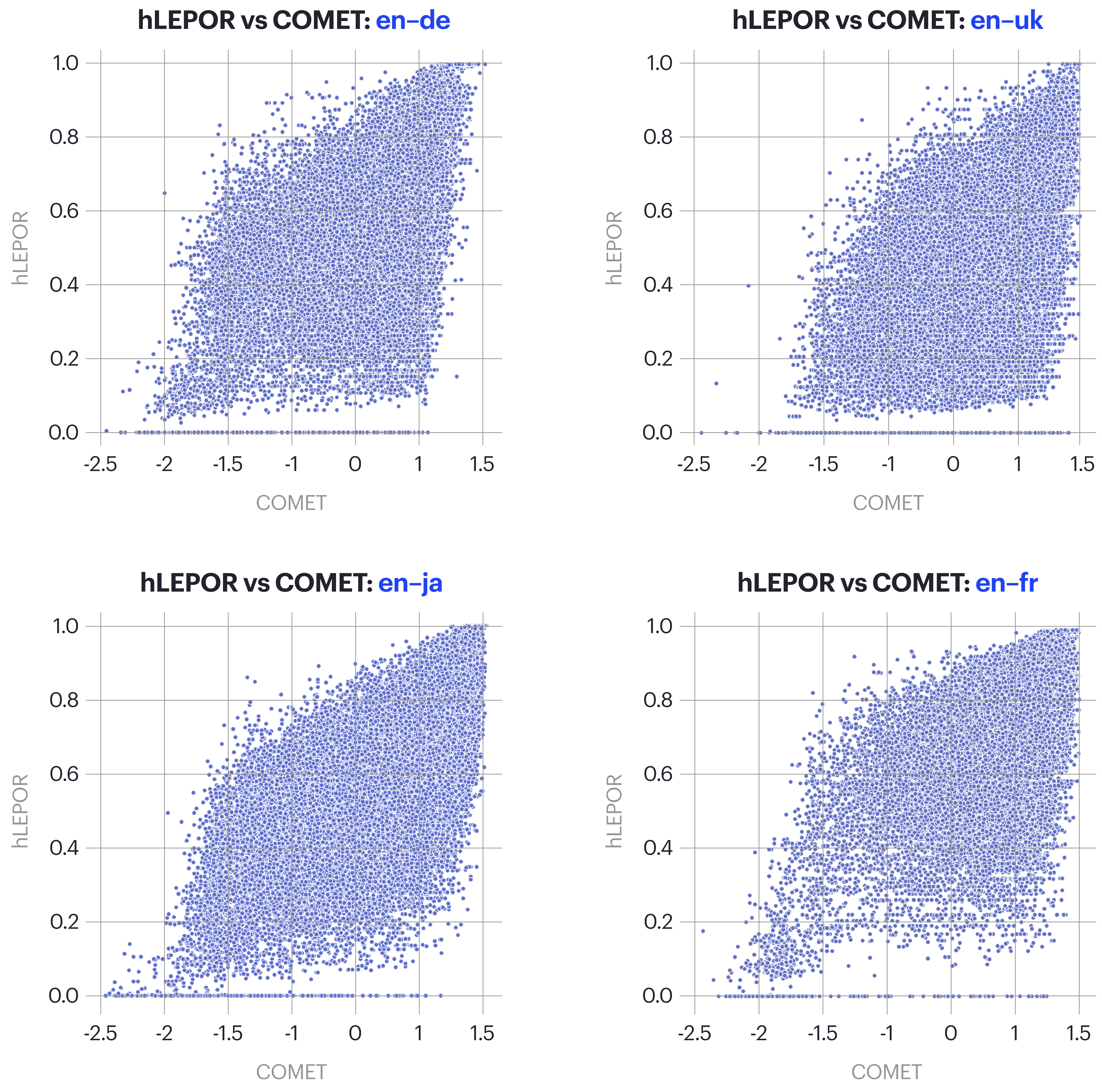
low hLEPOR + high COMET

→ paraphrases / synonyms

→ minor punctuation / tokenization issues

high hLEPOR + low COMET

→ COMET penilizes one-word omissions that do not affect hLEPOR that much



B.3 Comparing BERTScore and Prism

low BERTScore + high Prism

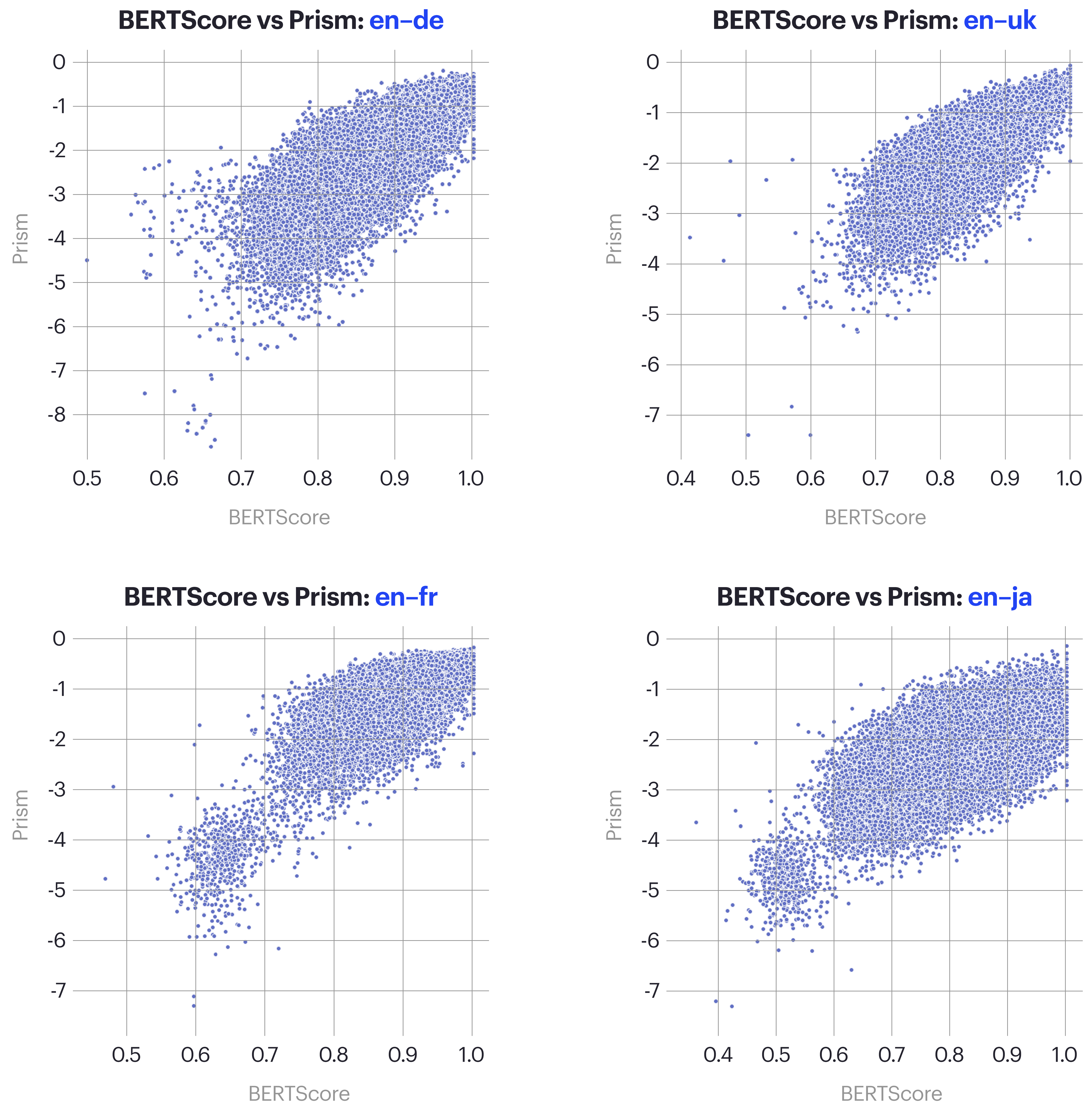
→ context-dependent alternative translations with different meanings (non-paraphrases)

→ non-translated phrases

→ punctuation issues that Prism does not penalize in some cases

high BERTScore + low Prism

→ PRISM for identical translations is not guaranteed to be close to 1 due to the logarithmic nature of the metric



B.4 Comparing COMET and Prism

low COMET + high Prism

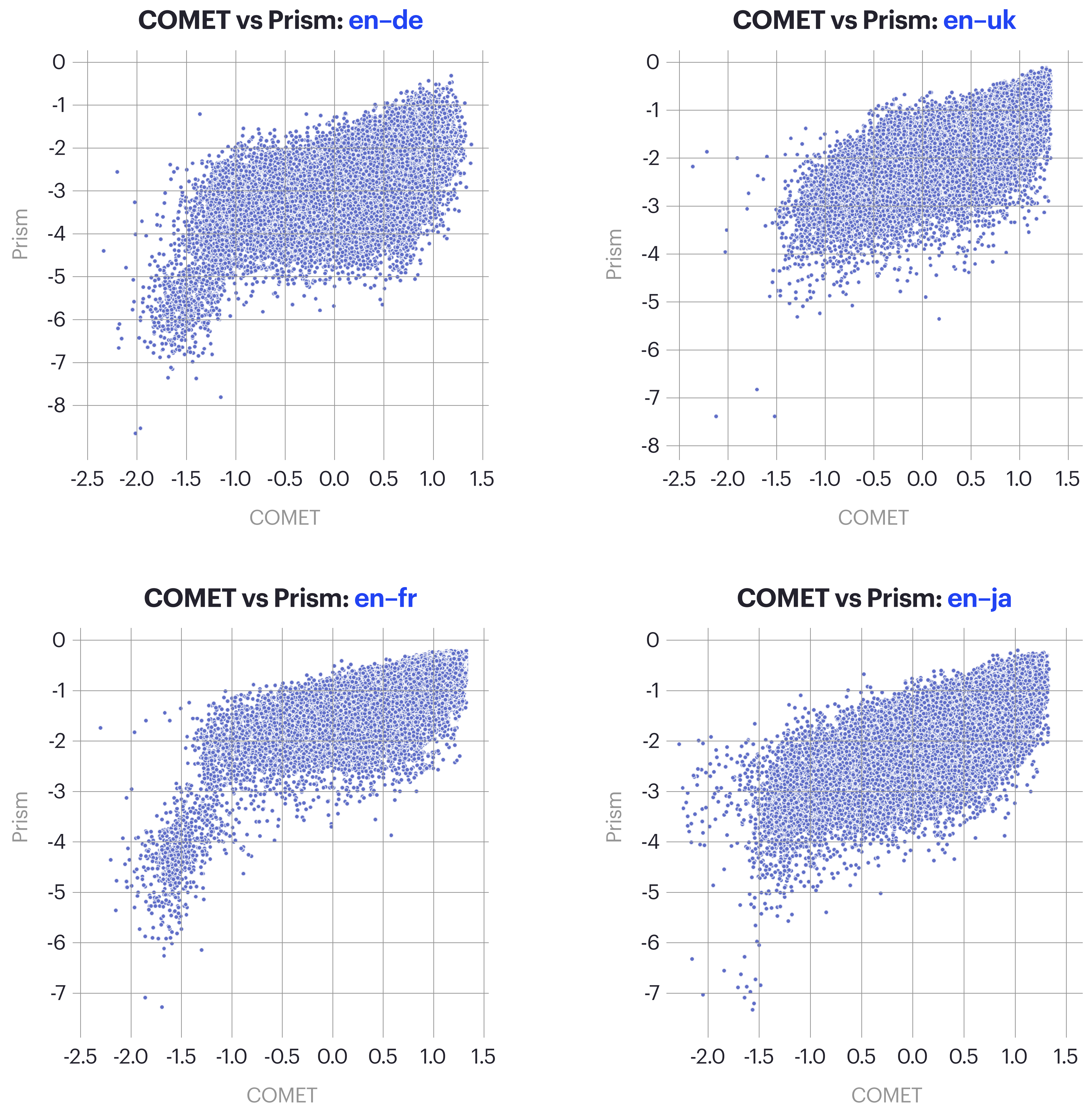
→ context-dependent alternative translations with different meanings (non-paraphrases)

→ punctuation issues that Prism does not penalize in some cases

high COMET + low Prism

→ PRISM for identical translations is not guaranteed to be close to 1 due to the logarithmic nature of the metric

→ punctuation issues that Prism penalizes too harshly in some cases



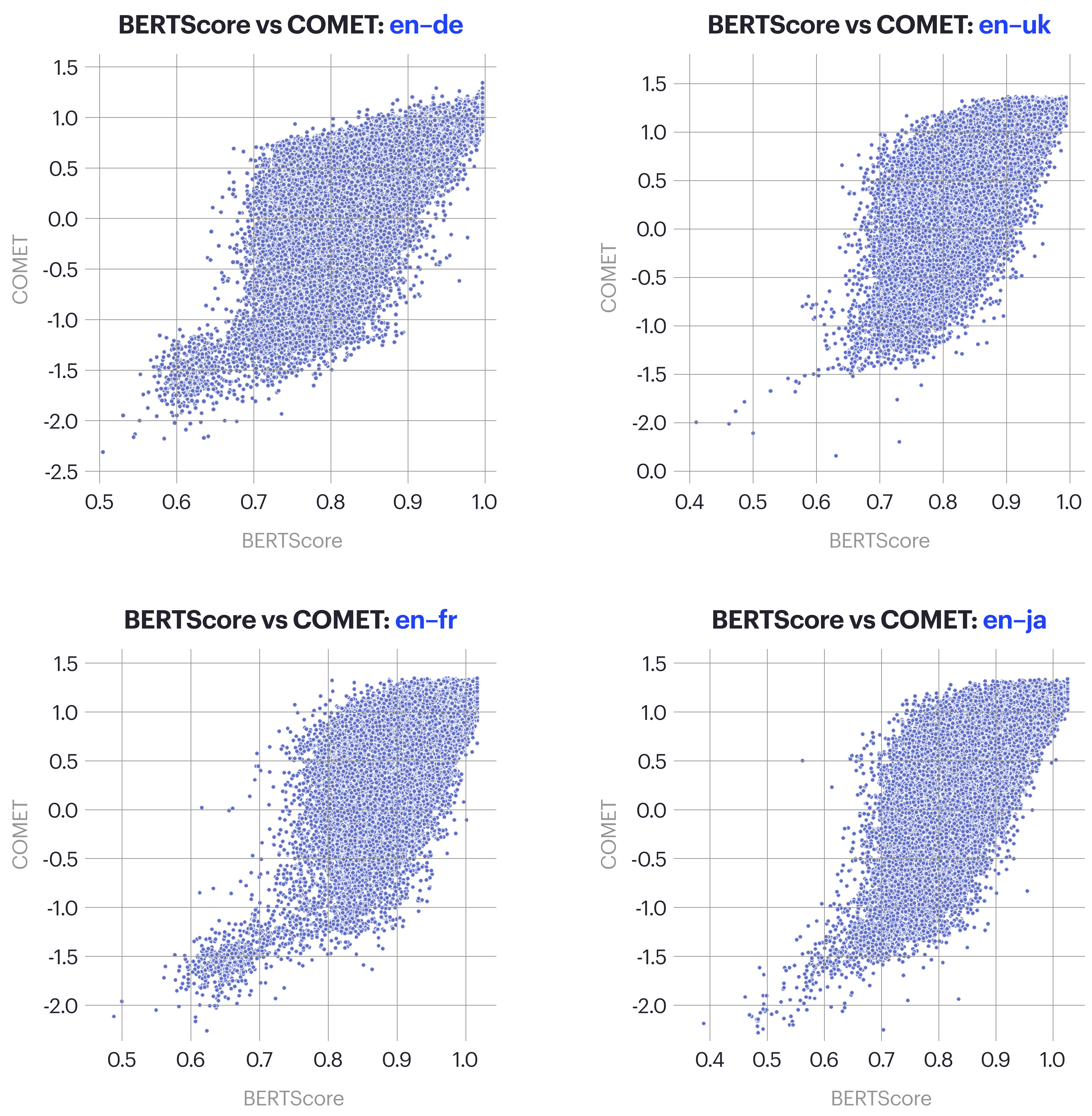
B.5 Comparing BERTScore and COMET

low BERTScore + high COMET

- → context-dependent alternative translations with different meanings (non-paraphrases)
- → minor tokenization issues (e.g. merging words vs using “-“ in German)

high BERTScore + low COMET

- → omissions and omissive paraphrases
- → context-dependent alternative translations with a different gender or tone of voice (mostly short sentences that lack context)



Appendix C

C.1 Ranking for BERTScore

→ For every language pair, we have normalized BERTScore to fit the [0,1] interval.

→ BERTScore rarely penalizes omissions and omissive paraphrases.

→ BERTScore penalizes different capitalization, therefore we have lowercased all text inputs. Per our observations, it does not lead to score corruption for properly capitalized sentences.

→ Does not reflect absolute quality level. Not comparable across language pairs.

MT vendors 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.

C.2 Best MT per Language Pair (BERTScore)

→ There are slightly more leading MT engines than COMET suggests, 8, with a similar amount of engines per language pair.

→ The same engines for minimal coverage: DeepL and Google.

→ Absolute values are not shown to avoid confusion, as the scores are not comparable across language pairs.

→ The domain and content type mix is different for every language pair (see the next slide) and greatly influences this leaderboard.

Best MT engines by normalized BERTScore

en–ar	Google			
en–de	DeepL	Google		
en–es	Amazon	Google	Microsoft	Yandex
en–fr	DeepL			
en–it	DeepL			
en–ja	DeepL			
en–ko	Google			
en–nl	DeepL			
en–pt	Google			
en–uk	Google			
en–zh	Baidu	Google	Youdao	

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

C.3 Best achievable score per Language pair and Domain (SacreBLEU)

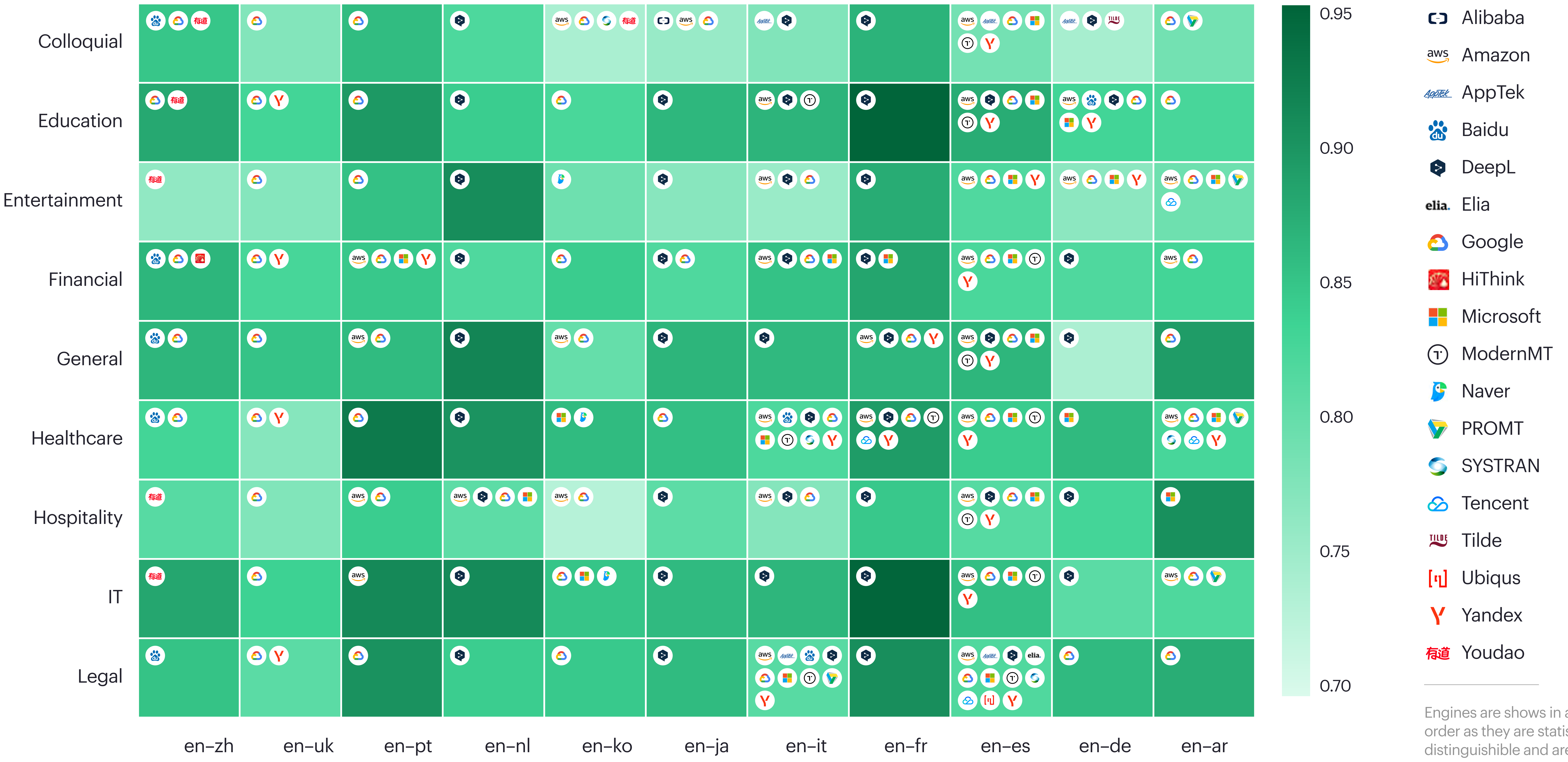
→ In the next slide, we show a heatmap of the best MT engines by a normalized COMET score. Each square shows the best providers for a particular language pair in a specific domain. The color of the square explains how high the best engines ranked among all engines in this combination of pair and domain.

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

→ Please remember that the absolute values of scores depend on the language pair you evaluate, and one should not compare scores between different language pairs.

MT vendors 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 MT engines by domain per normalized BERTScore



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

C.4 Best MT per Domain (BERTScore)

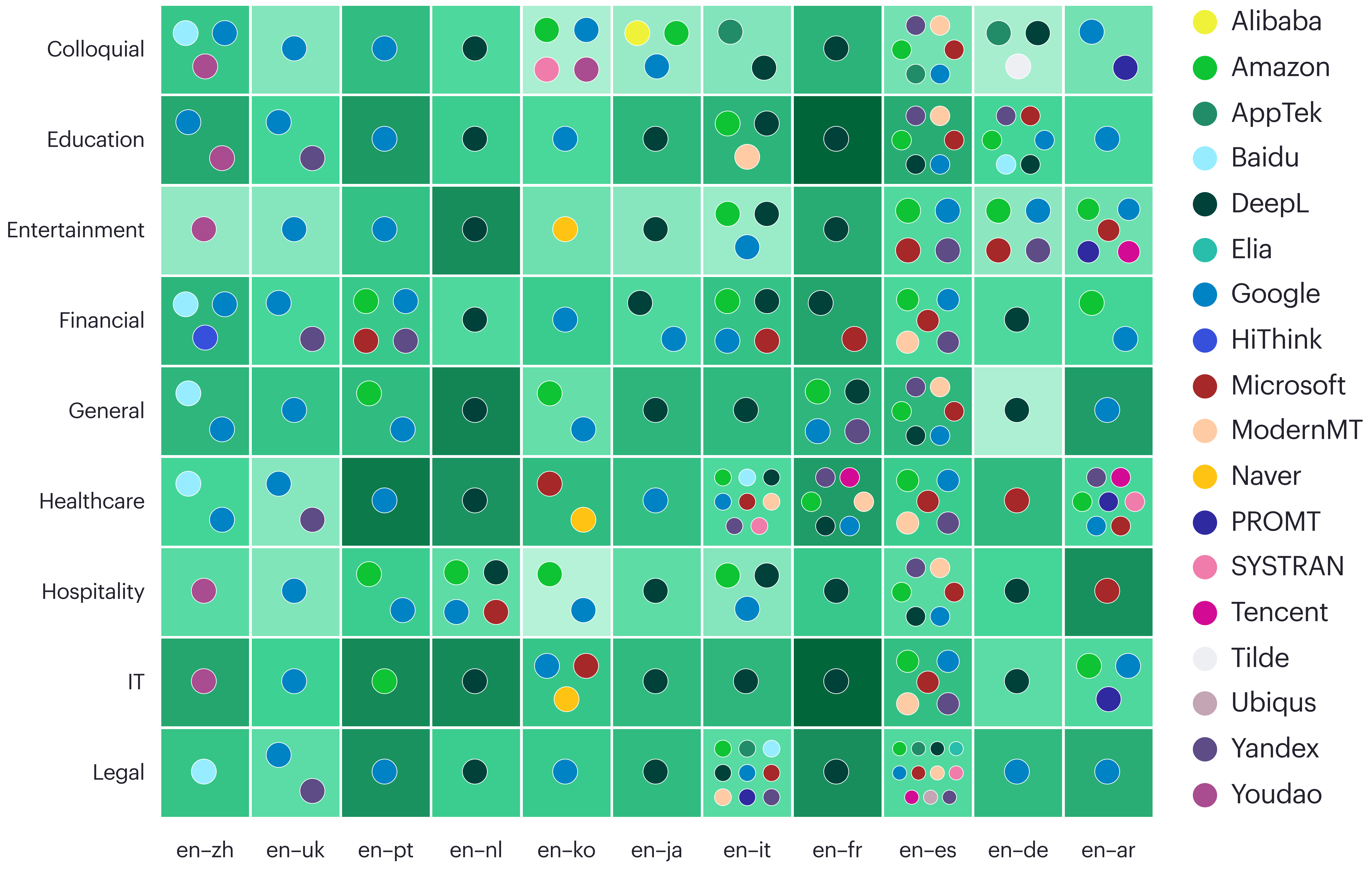
→ This chart is provided for reference. We recommend using the COMET chart on Slide 23.

→ 17 MT engines are among the statistically significant leaders for 9 domains and 11 language pairs.

→ The only significant difference from COMET is English to Chinese, Legal domain, where unlike COMET there is only one leading option, Baidu.

→ BERTScore favors Google a lot – our hypothesis is that because BERTScore is a Google product it might be more familiar with its translation style.

Available quality and best MT engines by domain per normalized BERTScore

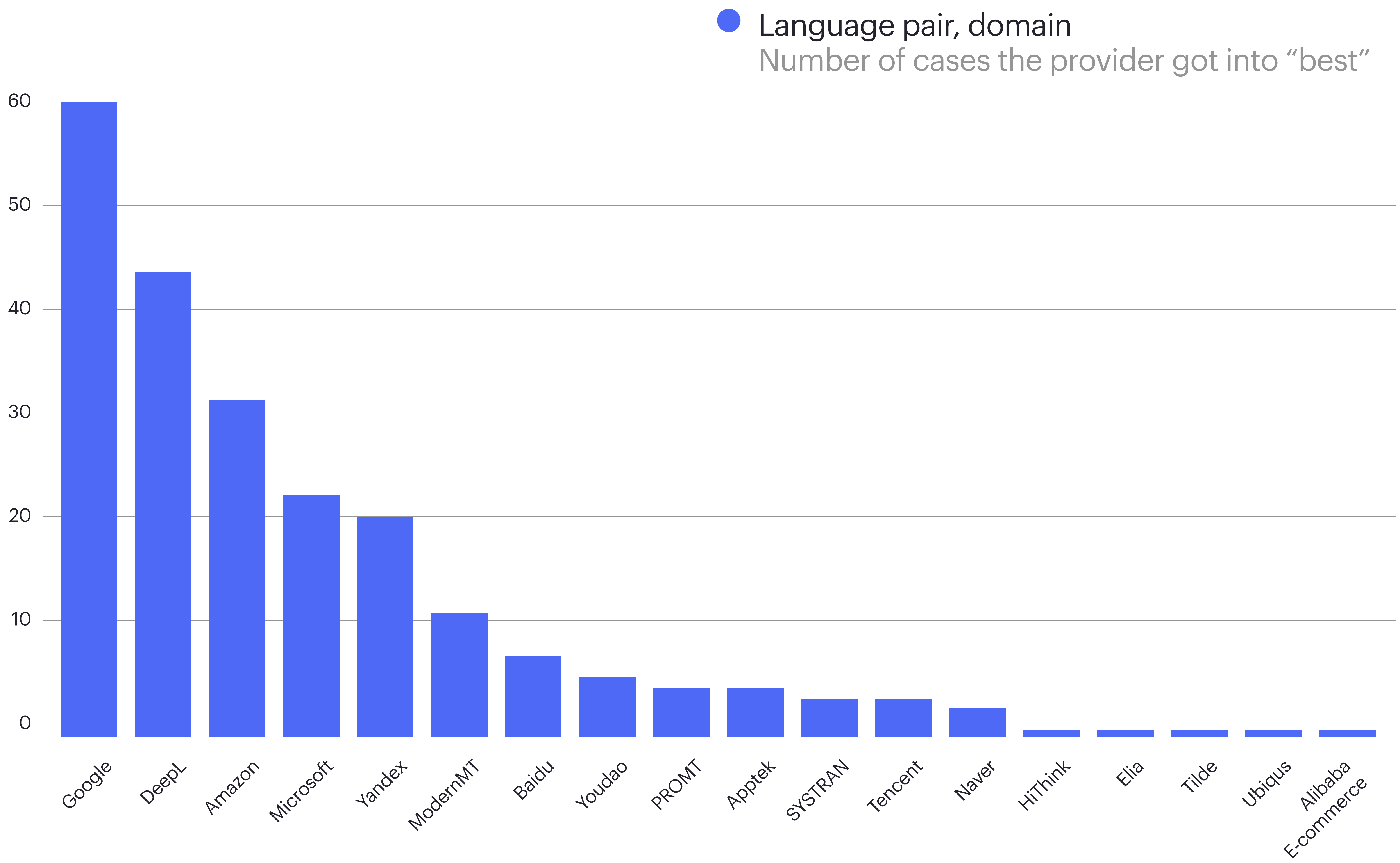


C.5 Top performing MT Providers (BERTScore)

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



Appendix D

D.1 Scores for Sentences of Different Lengths

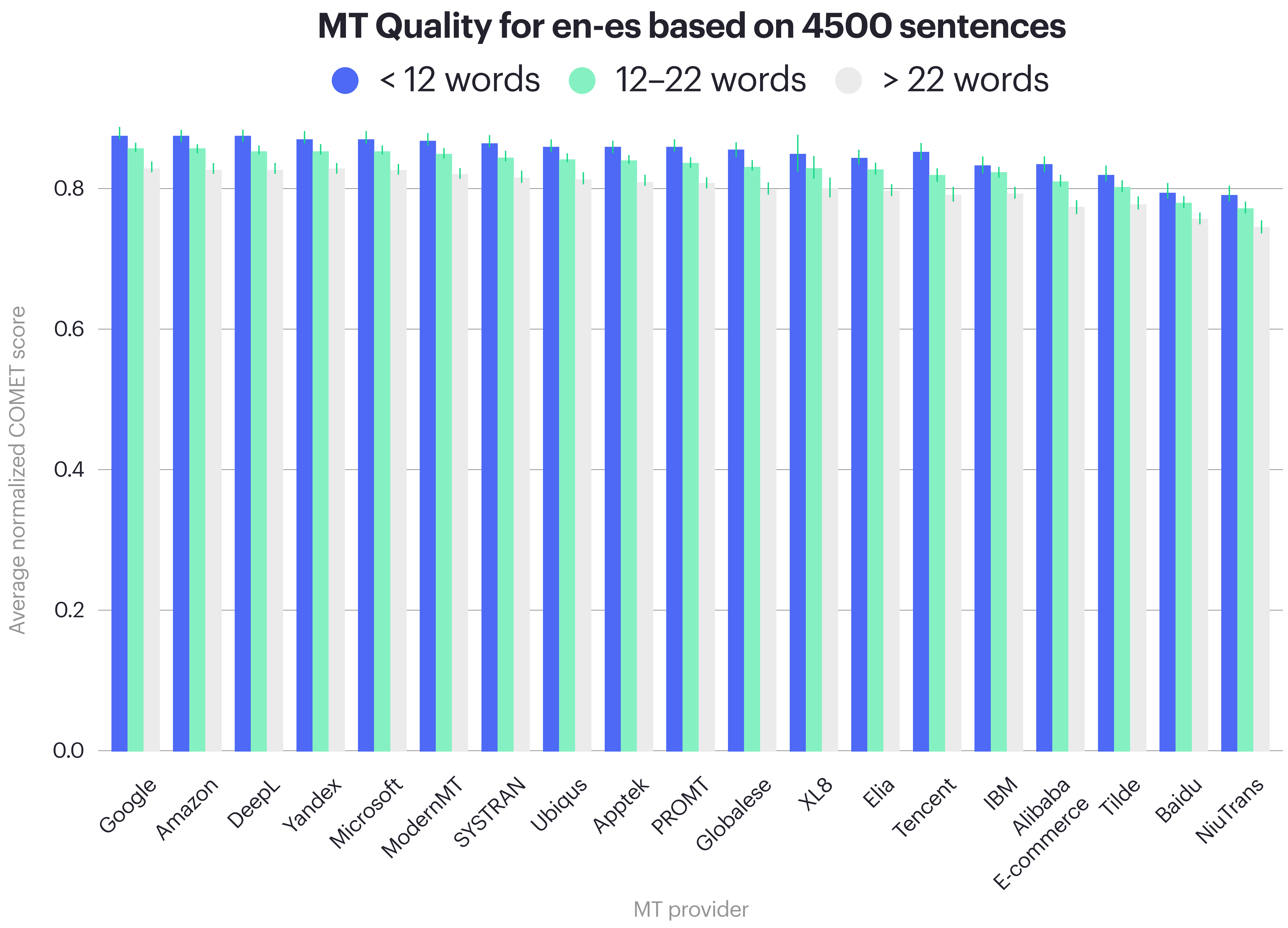
D.1 Scores for Sentences of Different Lengths

→ Typically, the scores are higher for shorter sentences.

→ English-to-Spanish demonstrates significant difference among MT engines for short and long segments (see the picture)

→ Some MT engines provide the top-tier scores for short and medium sentences, but fail to translate long ones, leading to the below average performance:

- Ubiqus for en-ja, en-uk
- Tencent for en-es
- Amazon for en-ar
- IBM for en-de



Appendix E

E.1 Best scores per domain (SacreBLEU)

E.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.

Highest SacreBLEU score for pair x domain

	en–ar	en–de	en–es	en–fr	en–it	en–ja	en–ko	en–nl	en–pt	en–uk	en–zh
Colloquial	21	29	45	67	36	27	18	50	69	26	56
Education	31	43	62	91	54	59	21	46	67	31	65
Entertainment	26	39	47	63	32	31	32	77	54	25	33
Financial	30	42	43	66	50	42	33	41	49	33	57
General	57	23	57	51	55	60	15	79	49	41	61
Healthcare	35	63	50	63	44	52	43	72	78	19	45
Hospitality	59	49	47	58	33	43	9	38	48	26	46
IT	34	40	54	89	59	59	41	77	72	39	67
Legal	51	63	44	73	40	59	33	50	72	31	56