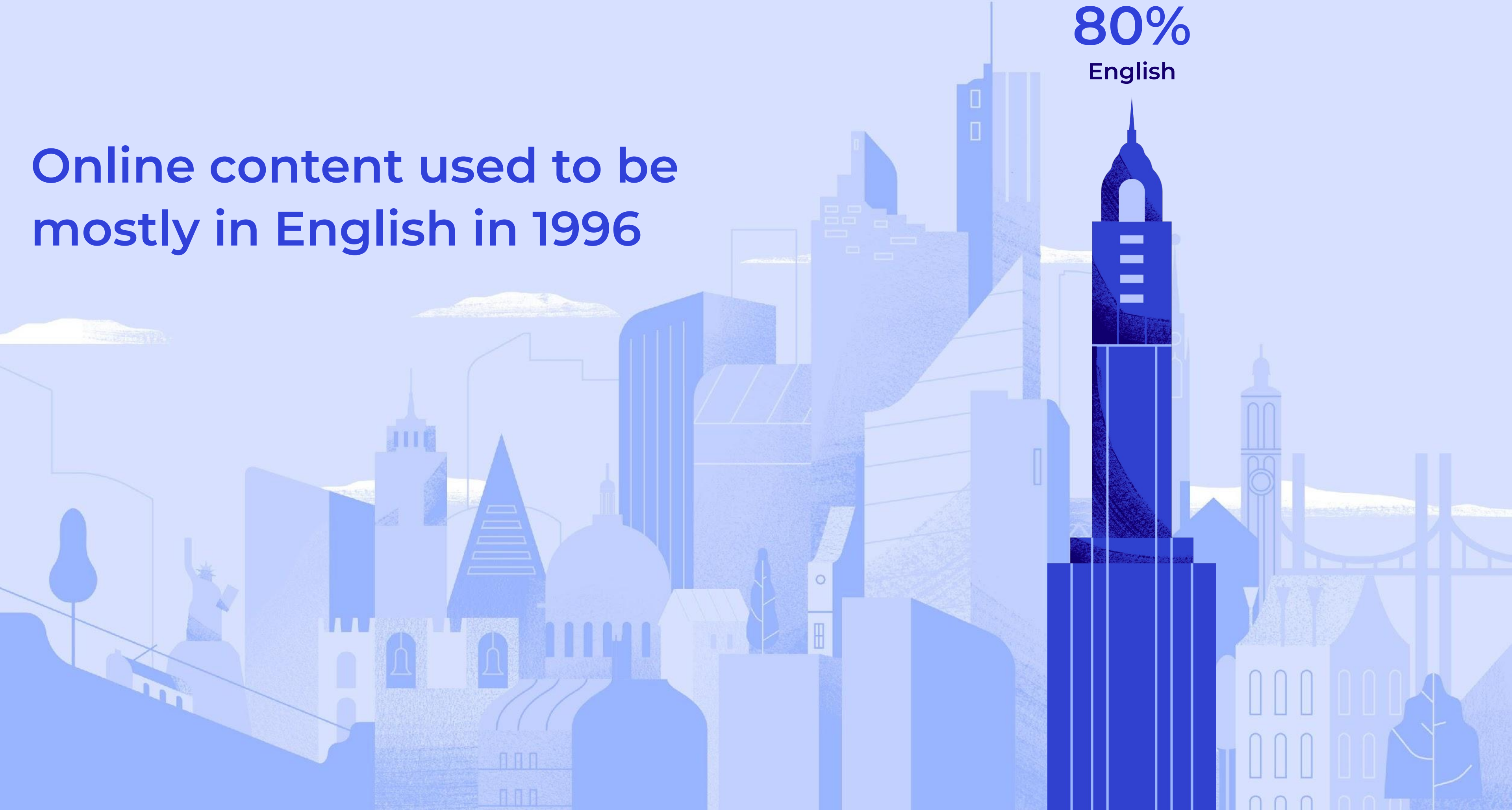
Unbabel

Tower LLM

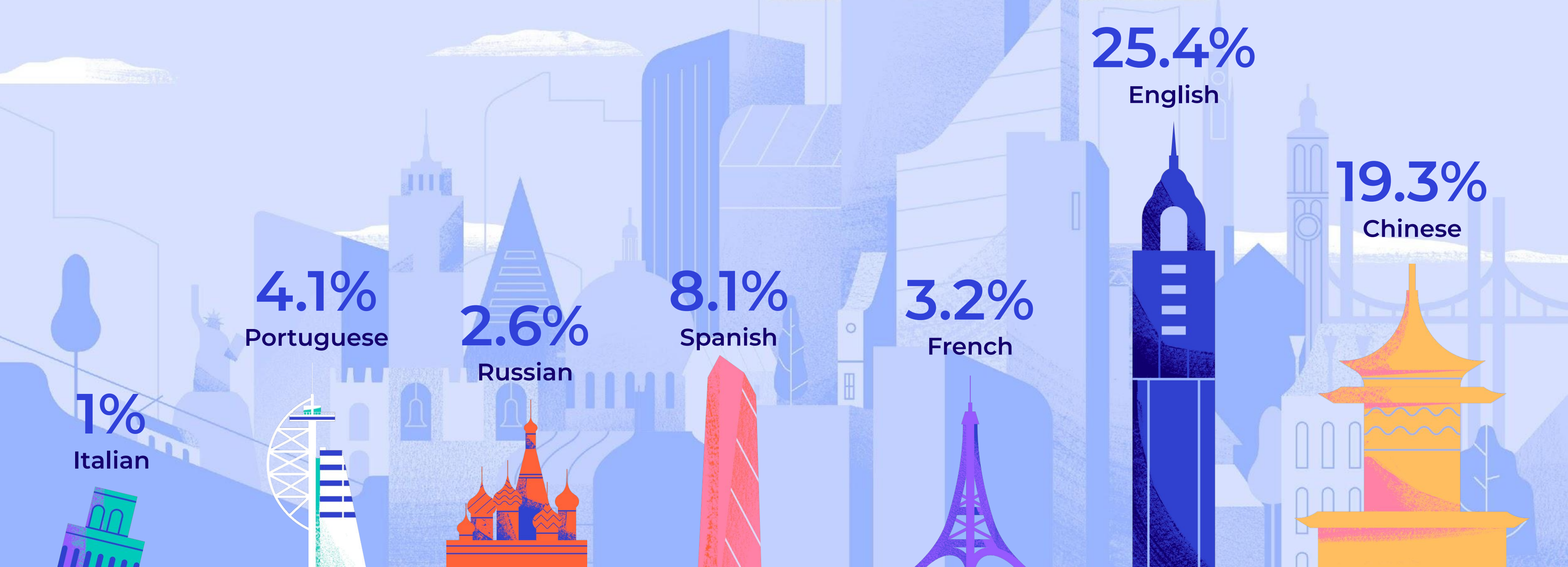
Large Language Models for Translation

Ricardo Rei

Online content used to be mostly in English in 1996



But the internet isn't in English anymore

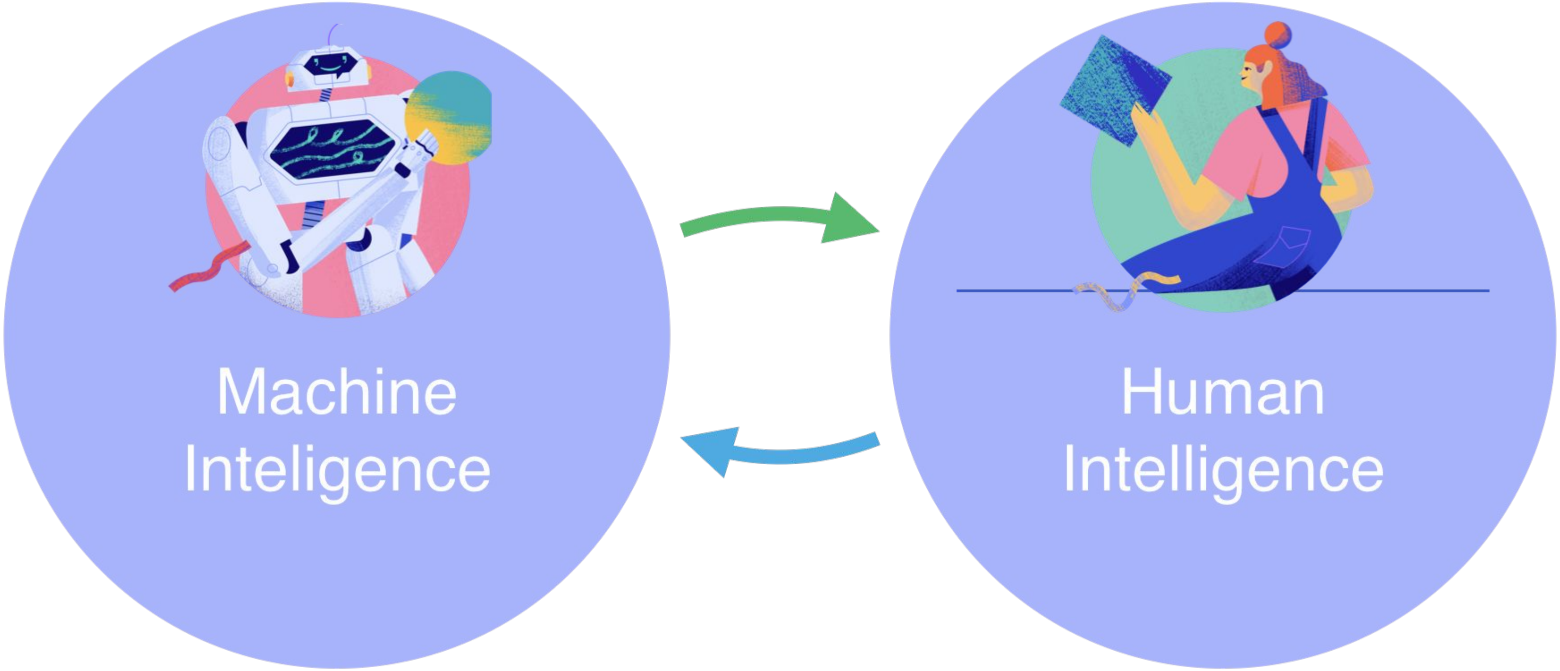


"All translation firms together are able to translate far less than 1% of relevant content produced everyday"

CSA – MT Is Unavoidable to Keep
Up with Content Volumes

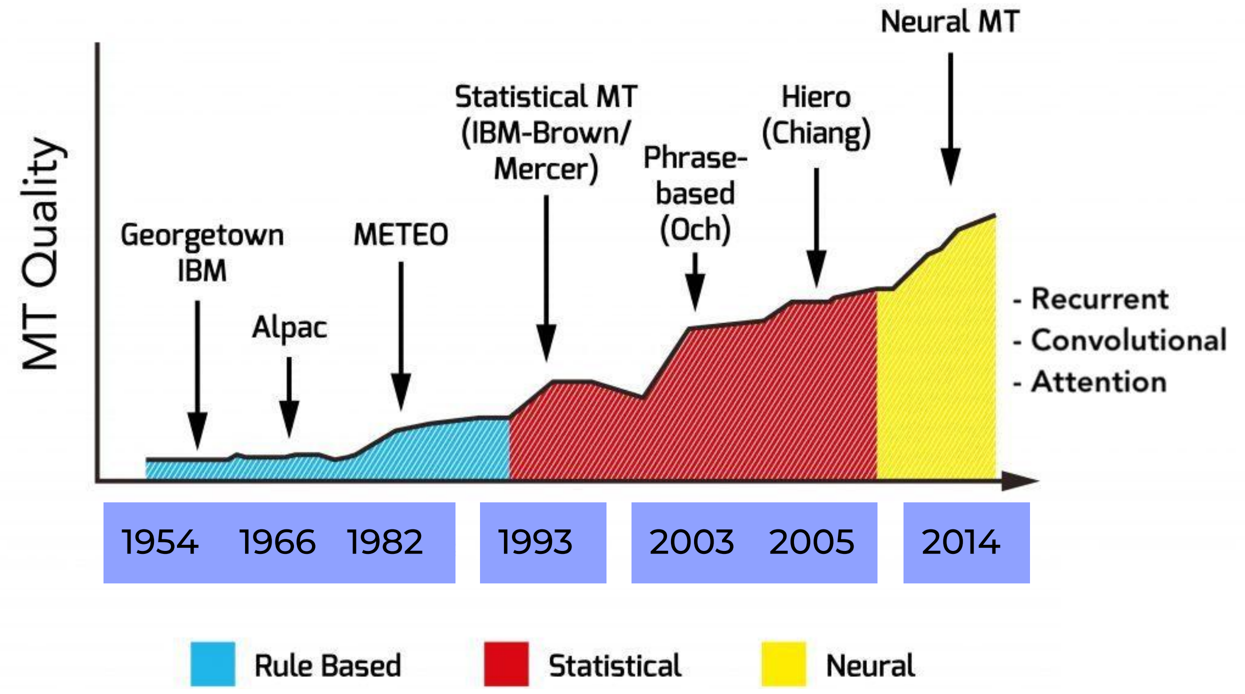
Tower LLM

Our View of the World

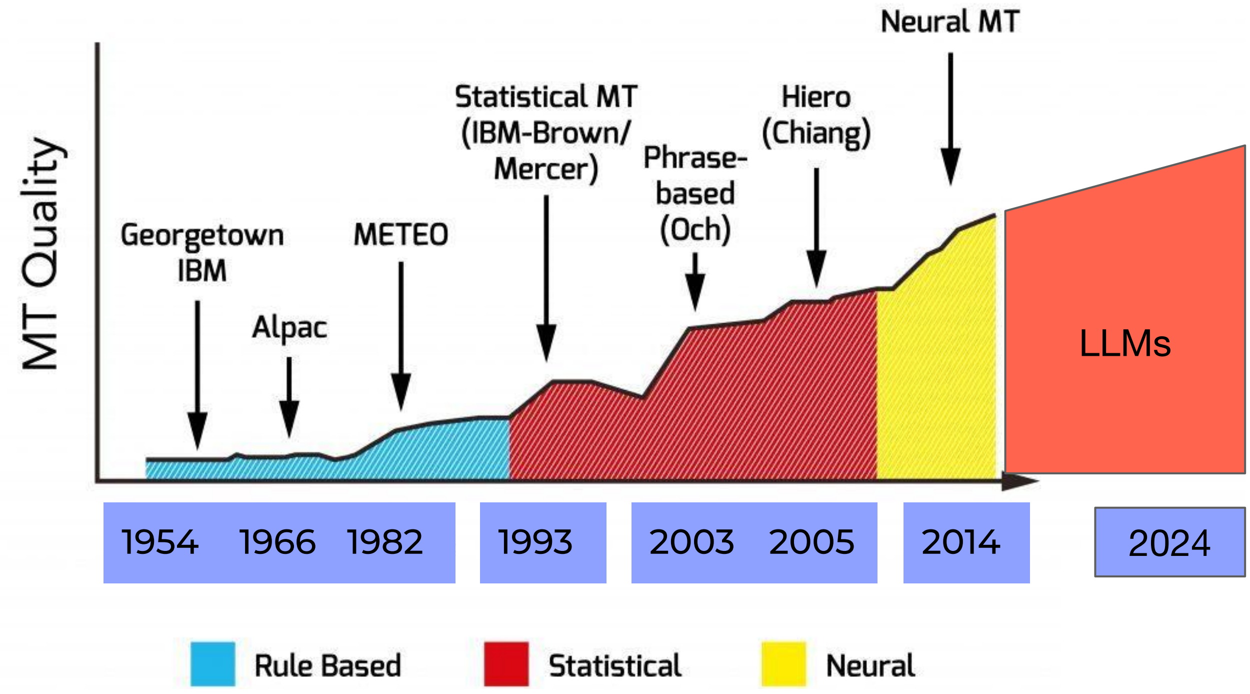


Unable to scale
to the growing mountains

History of Machine Translation – What's Next?

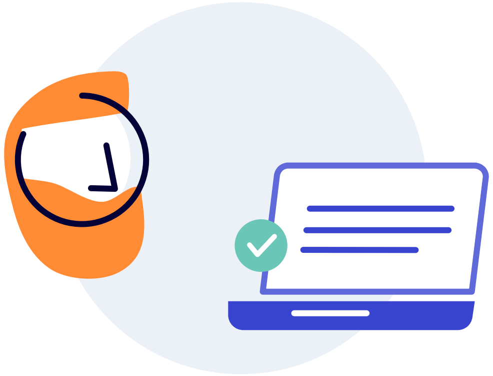


History of Machine Translation – What's Next?

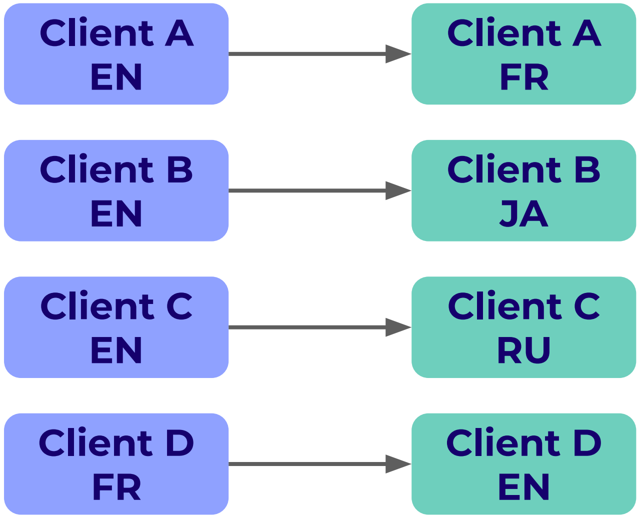


Some Open Problems with NMT systems

- Longer contexts (e.g. document-level MT, multilingual dialogue)
- How to use clients terminology?
- How to incorporate user translation style guide?
- High-risk translation (medical, legal, …)



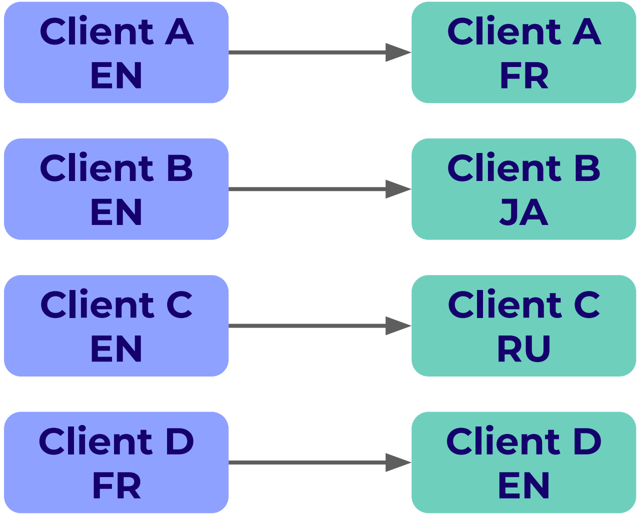
How do we adapt to style guide using NMT?



By training dedicated models for each client (most of the times bilingual models) the model learns to adapt to the clients language and terminology.

Tower LLM

How do we adapt to style guide using NMT?



By training dedicated models for each client (most of the times bilingual models) the model learns to adapt to the clients language and terminology.

Yet, this requires training and maintaining several models and it requires onboarding data!

Tower LLM

Towards LLM based MT

Glossary:
"February 14th" -> "2月14日"

Translate the source text from English to Chinese using the provided glossaries.

On February 14th, couples around the world celebrate Valentine's Day, expressing their affection with gifts and romantic gestures.

2月14日，世界各地的情 侣们庆祝情人节，用礼物和浪漫的举动表达他们的爱意。

Translate the following sentence into German:
TowerLLM can also do what normal NMT models do, but, due to prompting, it's much more flexible.

TowerLLM kann auch das, was normale NMT-Modelle können, aber aufgrund von Prompting ist es viel flexibler.

Consider the following translation rules:
Apple -> Apple
USD 1,000 -> 1.000 dólares
laptop -> computador portátil
buy -> adquirir

Translate the source text from English to Portuguese (Portugal) following the provided translation rules:

"The customer bought a last generation Apple MacBook laptop and paid USD 1,350."

"O cliente adquiriu um computador portátil Apple MacBook de última geração e pagou 1.350 dólares."

Unbabel

Tower LLM

Overview

Why Tower?

Unbabel

TÉCNICO LISBOA

université PARIS-SACLAY

Tower LLM

Tower is a big project

Amin Farajian

André Martins

Ben Peters

Duarte Alves

João Alves

José Pombal

José Souza

Manuel Faysse

Nuno Guerreiro

Patrick Fernandes

Pedro Martins

Pierre Colombo

Ricardo Rei

Sweta Agrawal

* Alphabetic order

The first suite of Tower models

Earlier this year we released the first version of Tower models that run at 7 and 13B params based on Llama 2.

TowerBase

Base model with **improved multilingual performance.**

TowerInstruct

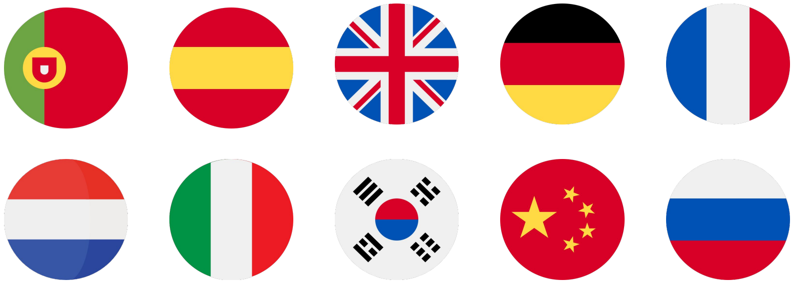
Optimized model
(built on top of TowerBase) for
translation-related tasks.

TowerLLM V1.0

Goal: create the best open multilingual LLM.

Focus (for now): **~10 (mostly) European languages**.

- The goal is **not** to go massively multilingual.



Tower LLM

TowerLLM V2.0

Goal: create the best open multilingual LLM.

Focus (for now): **15 (mostly) European languages**.

- The goal is **not** to go massively multilingual.

For WMT24 we added 5 more languages (Japanese, Hindi, Icelandic, Czech, Ukrainian)

We also replace Llama 2 models with Mistral and/or Llama 3.0

An LLM optimized for MT

TRANSLATION PIPELINE INSPIRED INSTRUCTION DATASET

Tasks that may feature in a classical production translation pipeline

Pre-translation

- Source correction
- Named-entity recognition
- Language identification

Translation

- General translation
- Translation modalities (context-aware, document-level, etc.)

Post-translation

- Automatic post-edition
- Error span prediction
- Critical error detection
- Translation quality ranking

MULTILINGUAL UNDERSTANDING

Tasks that may potentially help translation-related tasks by developing multilingual understanding

- Paraphrasing
- Word-sense disambiguation
- Cross-lingual summarization
- Multilingual QA

Tower LLM

TowerBase V2.0

From LLaMA-3 to TowerBase.

Llama 3

- ✓ Suite of models of different size
- ✓ A lot of open research on top of the models
- ✗ Not great for multilingual tasks

>

Extended multilingualization

How can we improve Llama 3 for multiple languages without compromising its general capabilities?

A Just instruction-tuning for the tasks of interest

B Continue pre-training on a large multilingual corpus (billions of tokens)

- **B1** Use only monolingual data 👍
- **B2** Mix monolingual and parallel data 👍

TowerBase V2.0

From LLaMA-3 to TowerBase.

Llama 3

- ✓ Suite of models of different size
- ✓ A lot of open research on top of the models
- ✗ Not great for multilingual tasks

>

Extended multilingualization

How can we improve Llama 3 for multiple languages without compromising its general capabilities?

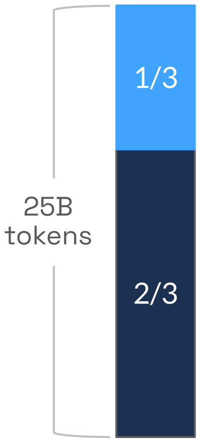
A Just instruction-tuning for the tasks of interest

B Continue pre-training on a large multilingual corpus (billions of tokens)

B1 Use only monolingual data

B2 Mix monolingual and parallel data

We built a corpus of 25B tokens with monolingual and parallel data



Parallel data

We used **OPUS** data all language pairs with English.
High quality filtering with **Bicleaner, CometKiwi-22**, etc.

Uniform weight across all language pairs.

In recent iterations we also prioritize paragraph/documents instead of short sentences

Monolingual data

We used curated monolingual data for all languages.
Filtering with **deduplication**, **language identification**, **perplexity**.
Uniform weight across languages.

Details on training TowerBase

Addition of parallel data

We append the parallel data as different documents of the format:

```
{SRC_LANG}: {SRC}\n{TGT_LANG}:
{TGT}<EOS>
```

Training Conditions

Single node of 8 x H100 GPUs for 7B
Multi node of 8 x H100 GPUs for 70B

Training Time

5/6 days for TowerBase 7B
1 week w/ 64 H100

TowerInstruct

From TowerBase to TowerInstruct.

TowerBase

- Multilingual capabilities
- Good few-shot performance
- No capability to follow instructions
- Suboptimal 0-shot performance

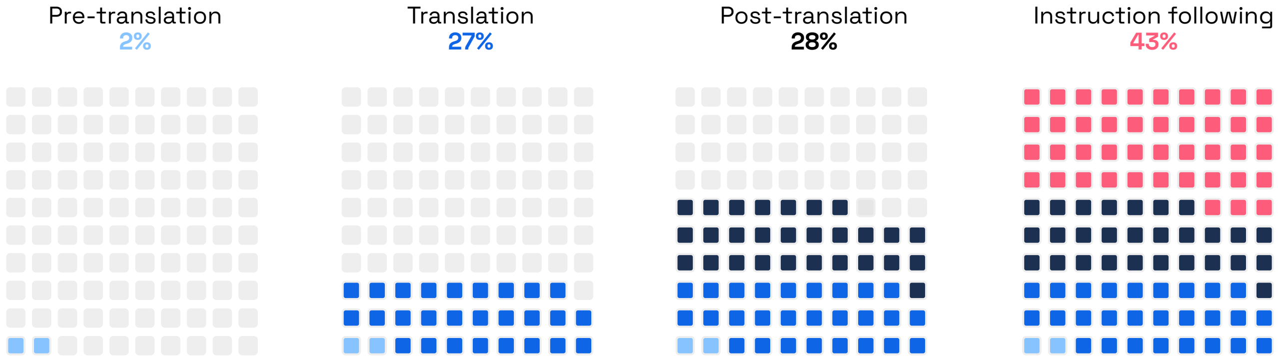
>

Instruction Tuning

How can we improve Tower's capabilities for tasks of interest? How can we make it a conversational model?

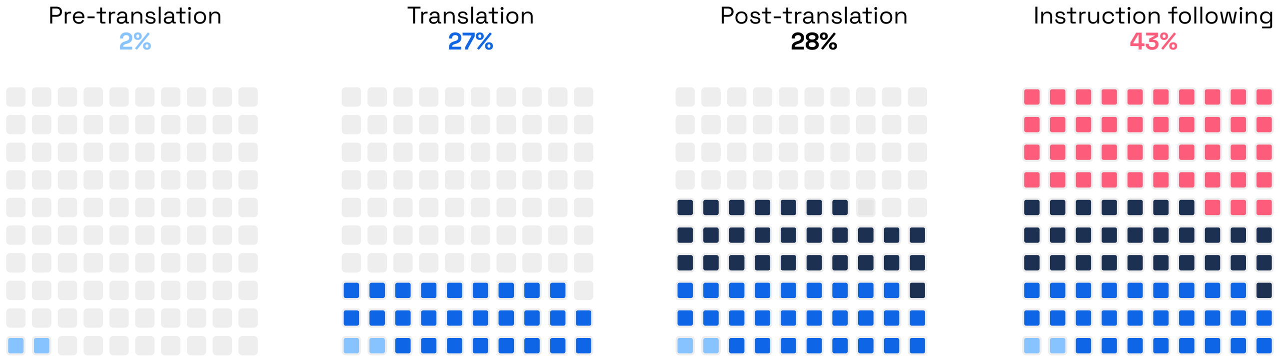
- **A** Collect lots of finetuning data and just train on that data
- **B** Collect fewer samples but guarantee they are high-quality
  - **B1** Use only finetuning data
  - **B2** Leverage conversational data and synthetic data from SOTA LLMs

Instruction following data:



Tower LLM

Instruction following data:

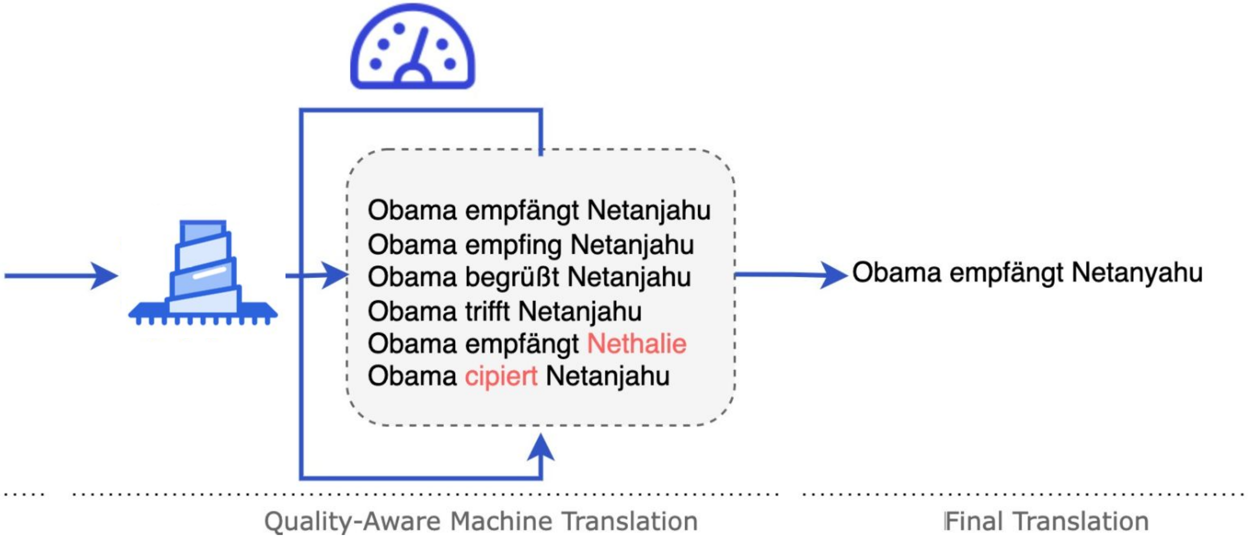


All data (specially the translation data) is highly curated!

Tower LLM

Recap on Quality-Aware decoding:

Translate the following sentence into German: **Obama receives Netanyahu**

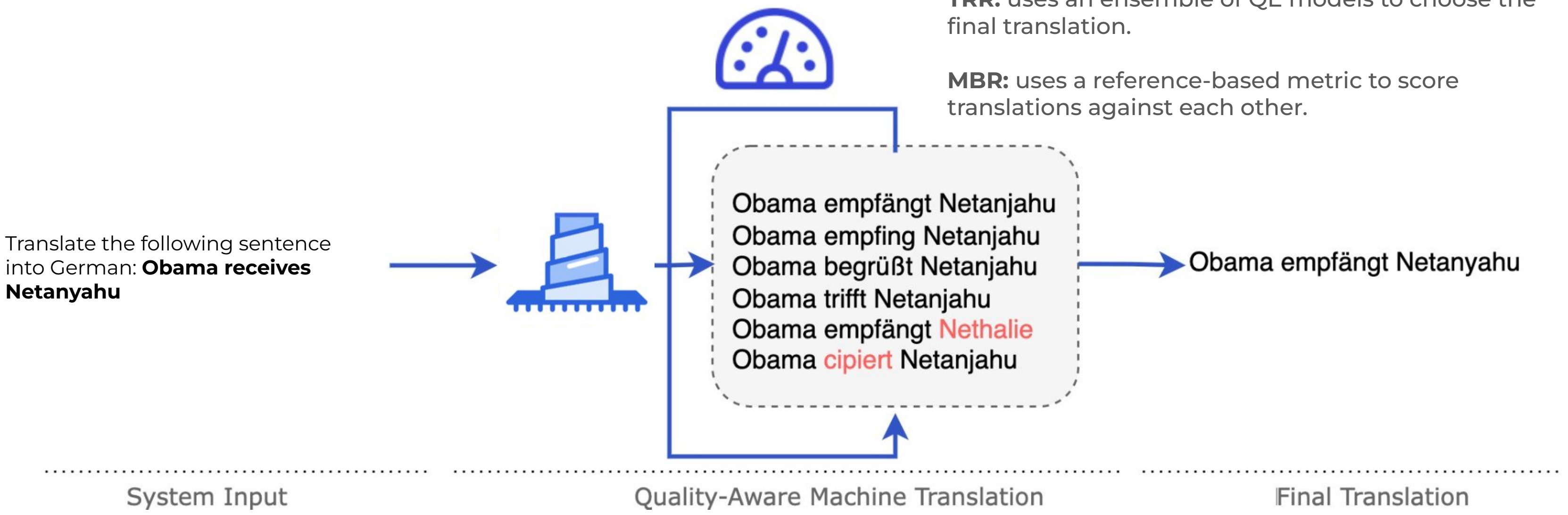


Tower LLM

Recap on Quality-Aware decoding:

TRR: uses an ensemble of QE models to choose the final translation.

MBR: uses a reference-based metric to score translations against each other.



Tower LLM

Results from Automatic Metrics

Models	en→xx METRICX↓	en→xx XCOMET↑	en→xx COMETKIWI↑	xx→yy METRICX↓	xx→yy XCOMET↑	xx→yy COMETKIWI↑
Baselines						
NLLB-54B	7.61 7	66.90 7	57.01 7	7.74 8	48.21 6	56.14 7
GPT-4O	1.50 6	83.74 6	77.04 5	2.18 5	70.44 2	76.19 4
CLAUDE-SONNET-3.5	1.40 5	84.85 5	78.09 4	1.98 4	69.73 2	76.77 4
DEEPL	—	—	—	4.38 6	56.19 4	68.33 6
TOWER						
TOWER-V2 7B	1.48 5	83.77 5	77.02 5	2.24 5	67.44 4	75.86 4
TOWER-V2 70B	1.32 4	84.87 4	78.29 4	2.04 4	69.20 3	76.70 4
TOWER + QAD						
TOWER-V2 70B+MBR	0.92 2	88.78 2	81.39 3	1.62 2	69.88 2	78.28 2
TOWER-V2 70B+TRR	1.03 3	87.95 3	82.13 2	1.73 2	**71.95** 1	79.38 2
TOWER-V2 70B 2-step	**0.89** 1	**89.25** 1	**82.54** 1	**1.58** 1	70.85 2	**79.69** 1

Table 2: Translation quality aggregated by language pairs on the WMT24 test set (without testsuites). We omit DEEPL from the en→xx averages because it does not support two language pairs. All metrics are their XXL variant.

Tower LLM

Results from Automatic Metrics

	en→xx			xx→yy		
Models	METRICX ↓	XCOMET ↑	COMETKIWI ↑	METRICX ↓	XCOMET ↑	COMETKIWI ↑
Baselines						
NLLB-54B	7.61 7	66.90 7	57.01 7	7.74 8	48.21 6	56.14 7
GPT-4o	1.50 6	83.74 6	77.04 5	2.18 5	70.44 2	76.19 4
CLAUDE-SONNET-3.5	1.40 5	84.85 5	78.09 4	1.98 4	69.73 2	76.77 4
DEEPL	—	—	—	4.38 6	56.19 4	68.33 6
TOWER						
TOWER-V2 7B	1.48 5	83.77 5	77.02 5	2.24 5	67.44 4	75.86 4
TOWER-V2 70B	1.32 4	84.87 4	78.29 4	2.04 4	69.20 3	76.70 4
AD						
70B+MBR	0.92 2	88.78 2	81.39 3	1.62 2	69.88 2	78.28 2
70B+TRR	1.03 3	87.95 3	82.13 2	1.73 2	**71.95** 1	79.38 2
70B 2-step	**0.89** 1	**89.25** 1	**82.54** 1	**1.58** 1	70.85 2	**79.69** 1

With Greedy decoding the Tower 70B is competitive to SOTA LLMs. The 7B model is not far behind

Table 2: Translation quality aggregated by language pairs on the WMT24 test set (without testsuites). We omit DEEPL from the en→xx averages because it does not support two language pairs. All metrics are their XXL variant.

Tower LLM

Results from Automatic Metrics

Models	en→xx			xx→yy		
	METRICX↓	XCOMET↑	COMETKIWI↑	METRICX↓	XCOMET↑	COMETKIWI↑
Baselines						
NLLB-54B	7.61 7	66.90 7	57.01 7	7.74 8	48.21 6	56.14 7
GPT-4O	1.50 6	83.74 6	77.04 5	2.18 5	70.44 2	76.19 4
CLAUDE-SONNET-3.5	1.40 5	84.85 5	78.09 4	1.98 4	69.73 2	76.77 4
	—	—	—	4.38 6	56.19 4	68.33 6
7B	1.48 5	83.77 5	77.02 5	2.24 5	67.44 4	75.86 4
70B	1.32 4	84.87 4	78.29 4	2.04 4	69.20 3	76.70 4
TOWER + QAD						
TOWER-V2 70B+MBR	0.92 2	88.78 2	81.39 3	1.62 2	69.88 2	78.28 2
TOWER-V2 70B+TRR	1.03 3	87.95 3	82.13 2	1.73 2	**71.95** 1	79.38 2
TOWER-V2 70B 2-step	**0.89** 1	**89.25** 1	**82.54** 1	**1.58** 1	70.85 2	**79.69** 1

Using quality-aware decoding methods like MBR we observe huge gains in automatic metrics

Table 2: Translation quality aggregated by language pairs on the WMT24 test set (without testsuites). We omit DEEPL from the en→xx averages because it does not support two language pairs. All metrics are their XXL variant.

Are we overfitting to Automatic Metrics?

To answer this question we have conducted human evaluation between the greedy outputs and the MBR/TRR outputs.

-Greedy is significantly worse.

- TRR and MBR are competitive. To get more concrete results we runned a second batch of evaluations using more “difficult” sentences. MBR was ranked higher than TRR.

Decoding	en→de	en→zh
Batch 1		
Greedy	85.43	84.11
TRR	87.16	85.55*
MBR	88.50*	85.47*
Batch 2		
TRR	—	68.55
MBR	—	72.76*

Table 3: SQM quality evaluation for three different decoding methods using TOWER-V2 70B. Numbers marked with an asterisk (*) are statistically significant. For English→Chinese, since the results of the first batch were not significant, we conducted a second batch comparison between TRR and MBR.

Tower LLM

Are we overfitting to Automatic Metrics?

To answer this question we have conducted human evaluation between the greedy outputs and the MBR/TRR outputs.

-Greedy is significantly worse.

- TRR and MBR are competitive. To get more concrete results we runned a second batch of evaluations using more “difficult” sentences. MBR was ranked higher than TRR.

We can trust that Greedy < MBR yet there might still be some bias when we compare to other models that do not use these metrics during inference

Decoding	en→de	en→zh
Batch 1		
Greedy	85.43	84.11
TRR	87.16	85.55*
MBR	88.50*	85.47*
Batch 2		
TRR	—	68.55
MBR	—	72.76*

Table 3: SQM quality evaluation for three different decoding methods using TOWER-V2 70B. Numbers marked with an asterisk (*) are statistically significant. For English→Chinese, since the results of the first batch were not significant, we conducted a second batch comparison between TRR and MBR.

Tower LLM

Are we overfitting to Automatic Metrics?

Let's look at MQM human evaluation:

En-De:

According to MQM GPT4 is ranked above Tower-70B: 1.649 < 1.683

According MetricX Tower-70B is ranked above GPT4: 1.1 < 1.4

According CometKiwi Tower-70B is ranked above GPT4: 72.2 > 70.1

En-Es:

According to MQM GPT4 is ranked above Tower-70B: 0.115 > 0.19

According MetricX Tower-70B is ranked above GPT4: 1.9 < 2.5

According CometKiwi Tower-70B is ranked above GPT4: 74.5 > 71.2

Are we overfitting to Automatic Metrics?

Let’s look at MQM human evaluation:

En-De:

According to MQM GPT4 is ranked above Tower-70B: 1.649 < 1.683

According MetricX Tower-70B is ranked above GPT4: **1.1 < 1.4**

According CometKiwi Tower-70B is ranked above GPT4: **72.2 > 70.1**

En-Es:

According to MQM GPT4 is ranked above Tower-70B: 0.115 > 0.19

According MetricX Tower-70B is ranked above GPT4: **1.9 < 2.5**

According CometKiwi Tower-70B is ranked above GPT4: **74.5 > 71.2**

Are we overfitting to Automatic Metrics?

Let's look at MQM human evaluation:

En-De:

According to MQM GPT4 is ranked above Tower-70B: **1.649 < 1.683**

According MetricX Tower-70B is ranked above GPT4: **1.1 < 1.4**

According CometKiwi Tower-70B is ranked above GPT4: **72.2 > 70.1**

En-Es:

According to MQM GPT4 is ranked above Tower-70B: **0.115 > 0.19**

According MetricX Tower-70B is ranked above GPT4: **1.9 < 2.5**

According CometKiwi Tower-70B is ranked above GPT4: **74.5 > 71.2**

According to automatic metrics Tower-70B is clearly outperforming other systems.

According to humans that is not the case. In most LPs Tower is a top-performing system but statistically tied with other best systems.

Impact of adding 5 languages

Two identical models:

- Model trained on 10 languages for 20B tokens
- Model trained on 15 languages for 25B tokens (20B for the initial 10 languages + 1B for each of the new langs)

Same SFT data with just a couple more translation samples added for the new languages (less than 5k samples)

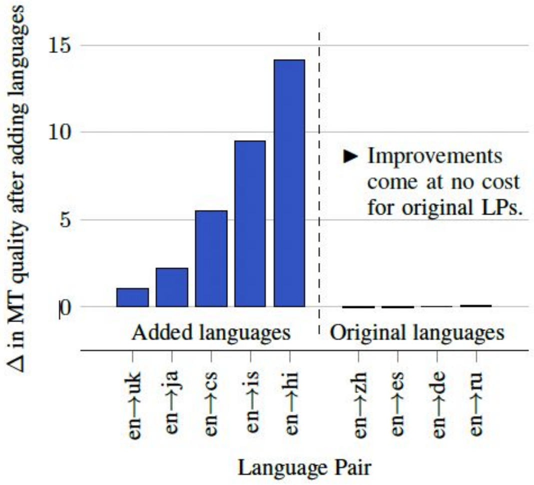


Figure 1: Improvement in MT quality after adding new languages to TOWER-V2; measured in negative METRICX-XXL-QE so taller bars equate to better quality.

Impact of adding 5 languages

Two identical models:

- Model trained on 10 languages for 20B tokens
- Model trained on 15 languages for 25B tokens (20B for the initial 10 languages + 1B for each of the new langs)

Same SFT data with just a couple more translation samples added for the new languages (less than 5k samples)

Can we keep increasing?

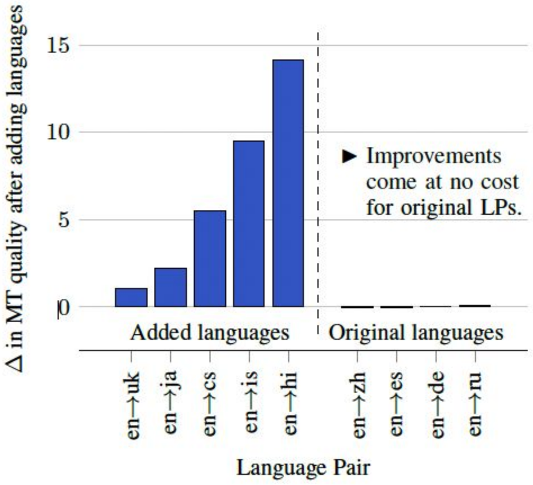


Figure 1: Improvement in MT quality after adding new languages to TOWER-V2; measured in negative METRICX-XXL-QE so taller bars equate to better quality.

Tower LLM

Impact of adding 5 languages

Two identical models:

- Model trained on 10 languages for 20B tokens
- Model trained on 15 languages for 25B tokens (20B for the initial 10 languages + 1B for each of the new langs)

Same SFT data with just a couple more translation samples added for the new languages (less than 5k samples)

Can we keep increasing?

Hard to answer... we are trying with up to 22 languages now.

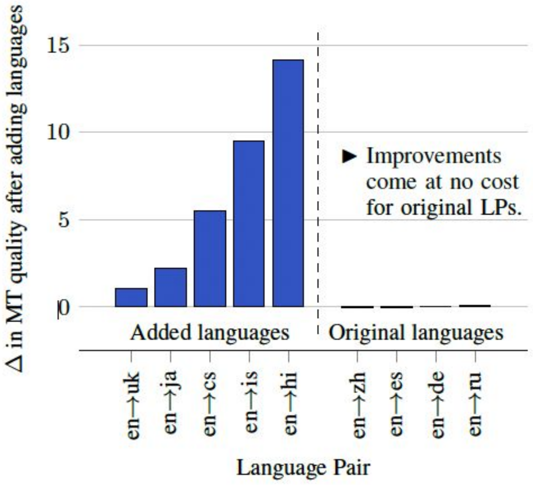
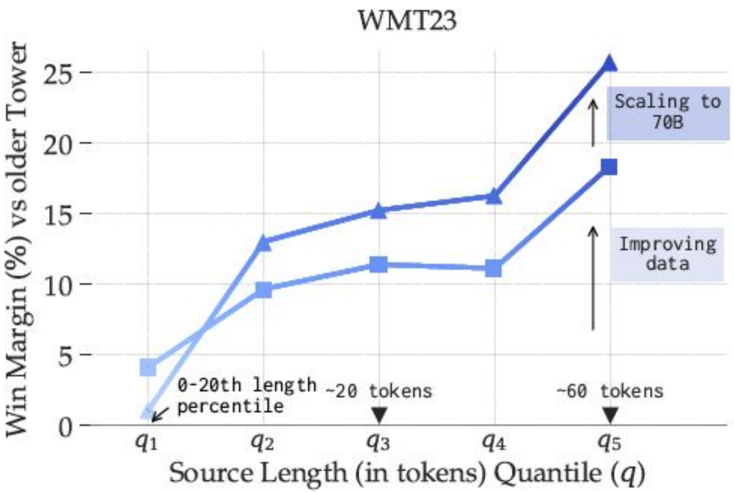
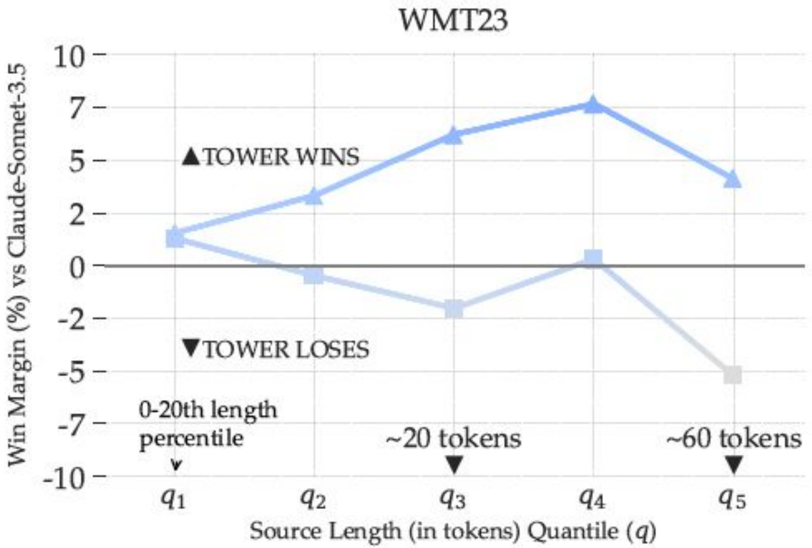


Figure 1: Improvement in MT quality after adding new languages to TOWER-V2; measured in negative METRICX-XXL-QE so taller bars equate to better quality.

Going beyond sentence-level MT





Going beyond sentence-level MT

Huge gains compared to version 1!

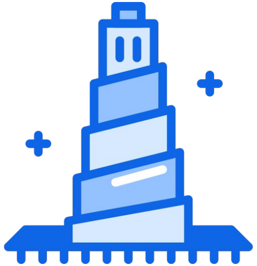
	WMT23-Paragraphs					
	en→xx			xx→yy		
Models	METRICX ↓	COMET ↑	CHRF ↑	METRICX ↓	COMET ↑	CHRF ↑
TOWER (older)	5.14	79.11	50.93	6.99	75.45	53.29
TOWER-V2-7B	2.72	84.45	54.35	1.87	87.57	61.36
TOWER-V2-70B	**2.40**	**84.87**	**55.06**	**1.72**	**87.75**	**62.29**

Context aware MT: Chat shared task

Translate the following sentence into German:
TowerLLM can also do what normal NMT models do, but, due to prompting, it's much more flexible.

Context:
USER: Gostaria de cancelar o meu Spotify
AGENT: I am so sorry hearing that you are interested in canceling your membership. To confirm your account cancellation please follow #URL#
USER: Desculpe mas não consigo abrir o link.
USER: A página não está disponível.

Translate the English source text to Portuguese, given the context.
English: No worries I can cancel it on my end.
Portuguese:



TowerLLM kann auch das, was normale NMT-Modelle können, aber aufgrund von Prompting ist es viel flexibler.

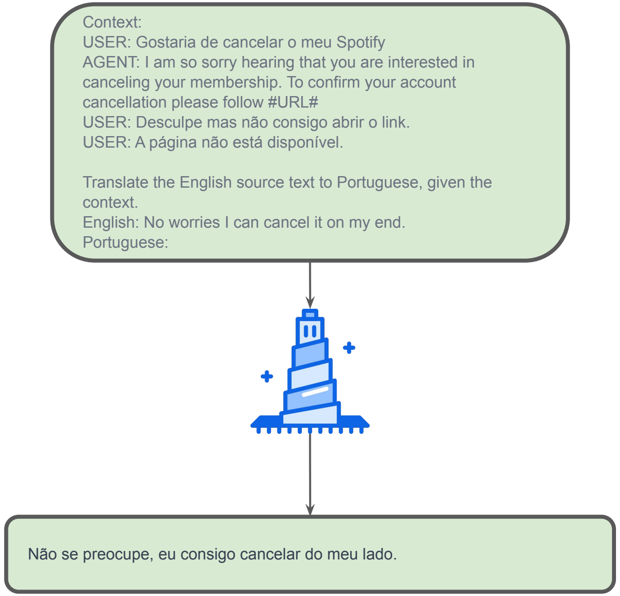
Não se preocupe, eu consigo cancelar do meu lado.

Context aware MT: Chat shared task

For few-shot we repeat the prompt 5 times.

Results are really high showing a great flexibility to “non standard” translation tasks.

Models	METRICX ↓	XCOMET ↑
	en→xx	
TOWER-V2-70B 0-shot	0.510	96.96
TOWER-V2-70B 5-shot	0.495	96.89
	xx→en	
TOWER-V2-70B 0-shot	1.051	94.84
TOWER-V2-70B 5-shot	0.766	95.54



A lots of ongoing work

- Analysis
- Alignment
- New Tasks
- Decoding Strategies (potential biases)
- Scaling Up and MoEs

Next steps: on the road to EuroLLM...

EuroLLM-1.7B model trained from scratch

- A **1.7B model trained from scratch on 4T tokens** on 35 languages:
  - Support for all 24 official EU languages + strategic languages (e.g. Chinese, Russian, etc)
  - Includes parallel data from the pre training phase (similar to Palm 2)
  - Developed several scaling laws to predict the performance of the 1B model;
  - Competitive to Gemma 2B but highly multilingual

EuroLLM 9B is at 50% of its total training (4T tokens) and it already shows better MT results than Gemma 9B and Llama 3.1 8B

Tower LLM

DEMO