GenAI for Translators: An Introduction to LLMs and their Usage

Antonio Toral

#TQ26: Traduction & Qualité 2026

Université de Lille, 30/01/2026

UA Universitat d'Alacant
Universidad de Alicante

LT-LiDER

LT-LiDER (ref. 2023-1-ES01-KA220-HED-000161531) has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the European Union

The consortium

UAB
Universitat Autònoma de Barcelona

DCU

Technology Arts Sciences
TH Köln

universität wien

UGA
Université Grenoble Alpes

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

rijksuniversiteit groningen

LT-LiDER

LT-LiDER (ref. 2023-1-ES01-KA220-HED-000161531) has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the European Union

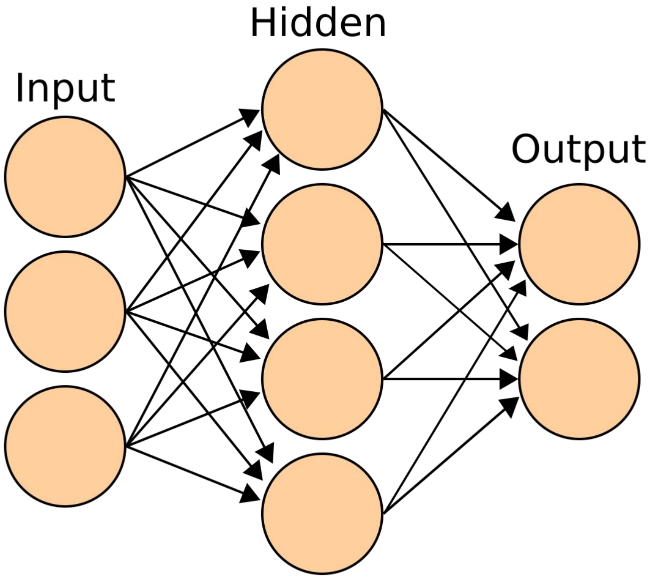
Contents

1. NMT and LLMs 101. What is different?
2. What is the current SOTA in MT?
3. Can MT be creative? And natural?
4. Why and how LLMs in your own computer?

1. LLMs 101

“Les outils d’IA générative quant à eux font une apparition remarquée avec une utilisation présente chez 34% des agences et 43% des indépendants.”

NMT and LLMs 101



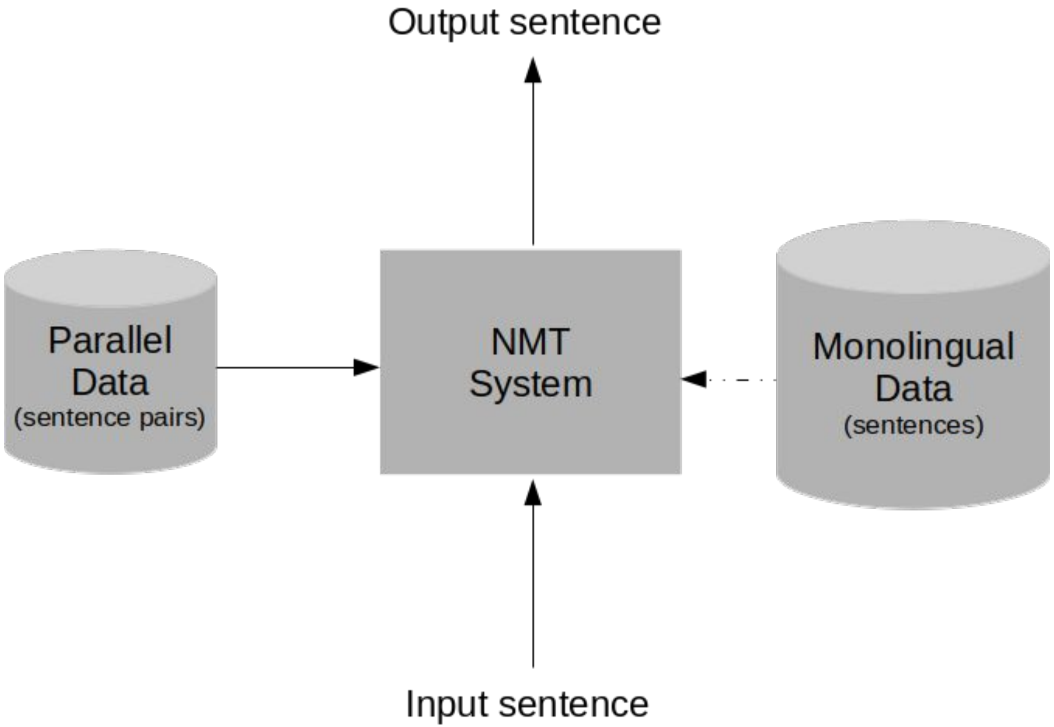
Source: Wikipedia (GFDL)

- 3 layers
  - 1 input
  - 1 hidden
  - 1 output

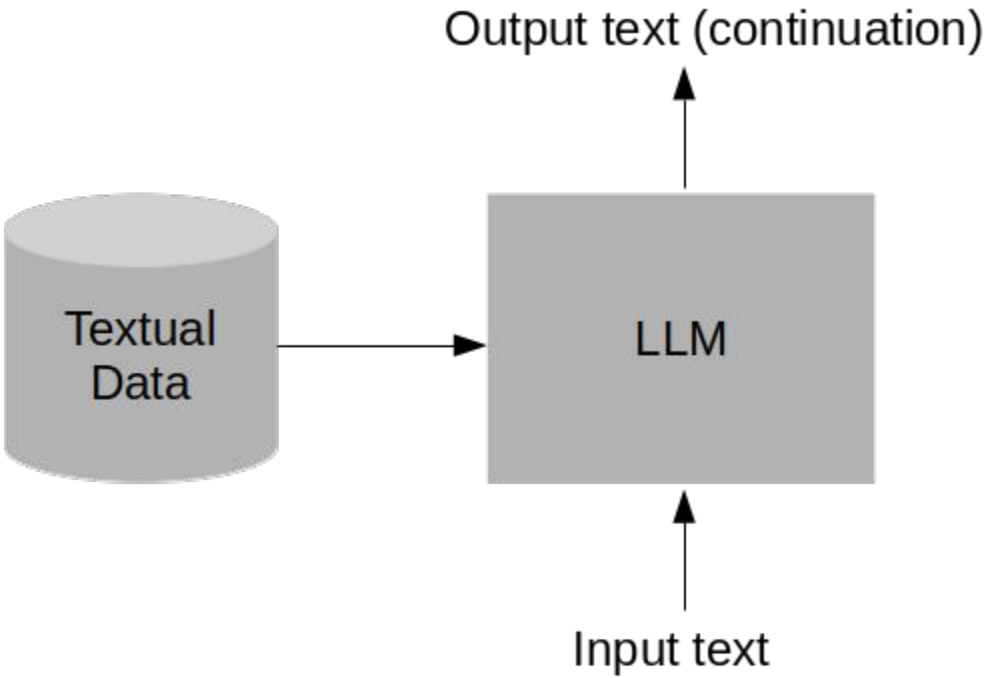
- 20 parameters

GPT-4: 1.76 trillion parameters (estimated)
1,760,000,000,000

NMT: predict the translation



LLM: predict the next word



Output text (continuation)

Textual Data

LLM

Input text

LLM Types

- **Base**. Trained to predict the next word in a text

LLM Types

- **Base**. Trained to predict the next word in a text
- **Instruct**. Post trained with some tasks (instructions)
- **With “reasoning”**. Post trained to reply in two parts
    1. “reasoning” about the problem
    2. answer

Translations by LLMs vs NMT

✅

- Beyond simply translating
- Use of global context
- Better quality for languages with many resources
- Less literal translations (Raunak et al., 2023)

❌

- Higher computational cost

2. SOTA MT

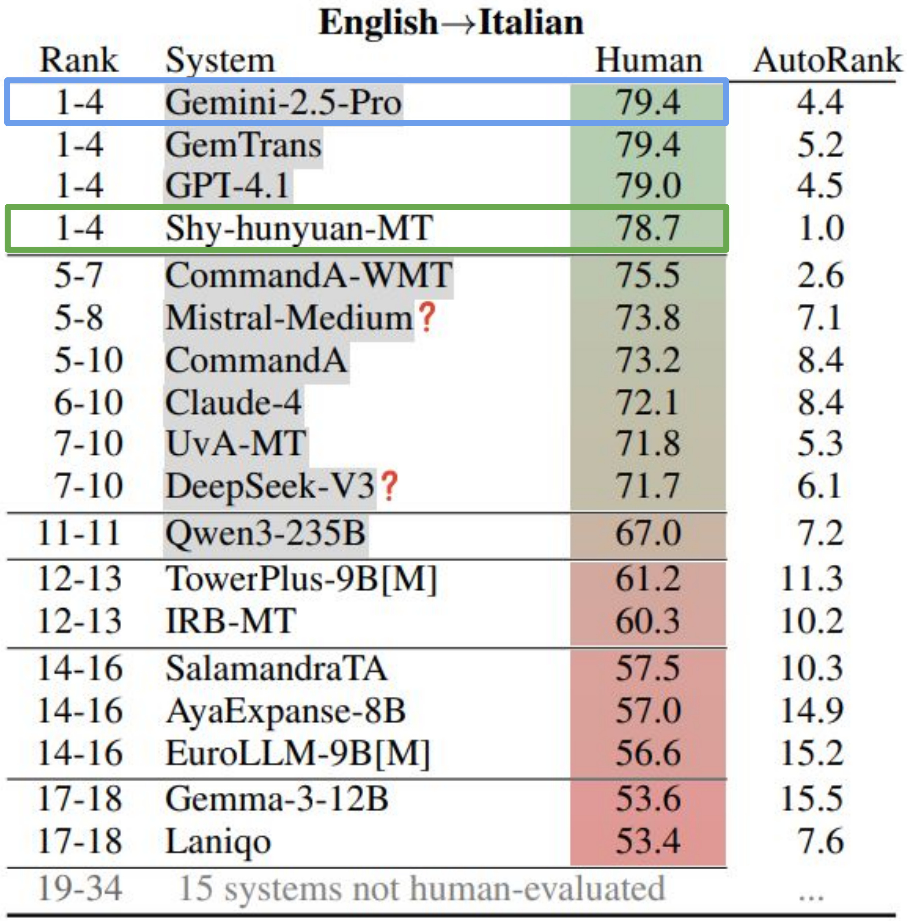
“la TA [est utilisé] par environ 1 professionnel sur 2. Les outils d’IA générative quant à eux font une apparition remarquée avec une utilisation présente chez 34% des agences et 43% des indépendants.”

Research: WMT 2025

- Best system: a general-purpose LLM
  - Gemini 2.5 Pro (‘thinking’ mode)

- Good results from ‘small’ LLMs optimised for MT
  - Effective techniques: distillation, reinforcement learning, etc.
  - Models: Shy-hunyuan-MT (7B) (Zheng et al. 2025), Algharb (14B) (Wang et al. 2025)

Human eval @ WMT’25



English→Italian

Rank	System	Human	AutoRank
1-4	Gemini-2.5-Pro	79.4	4.4
1-4	GemTrans	79.4	5.2
1-4	GPT-4.1	79.0	4.5
1-4	Shy-hunyuan-MT	78.7	1.0
5-7	CommandA-WMT	75.5	2.6
5-8	Mistral-Medium?	73.8	7.1
5-10	CommandA	73.2	8.4
6-10	Claude-4	72.1	8.4
7-10	UvA-MT	71.8	5.3
7-10	DeepSeek-V3?	71.7	6.1
11-11	Qwen3-235B	67.0	7.2
12-13	TowerPlus-9B[M]	61.2	11.3
12-13	IRB-MT	60.3	10.2
14-16	SalamandraTA	57.5	10.3
14-16	AyaExpanse-8B	57.0	14.9
14-16	EuroLLM-9B[M]	56.6	15.2
17-18	Gemma-3-12B	53.6	15.5
17-18	Laniqo	53.4	7.6
19-34	15 systems not human-evaluated		...

Industry (Pangeanic, 2025)

Disadvantages of LLMs: hallucinations, terminology handling, high computational cost

Recommendations

- **NMT** for large volumes of data, with specific terminology, and privacy-sensitive
- **LLMs** for creative, narrative, and marketing texts
- **Hybrid** (NMT and LLMs): NMT-level control with LLM-level fluency

3.1. Creative MT?

“les textes générés automatiquement peuvent-ils être aussi créatifs que des textes rédigés par des professionnel(le)s de la traduction ?”

3.1. Creative MT?

Creativity requires both **originality** and **effectiveness.**

Runco and Jaeger 2012, 93

https://esp.grandado.com/

How to Annotate Creativity? Novelty

Bayer-Hohenwarter (2009, 2011)

Everything
was perfectly swell.

Unit of
creative
potential
(UCP)

Todo
era absolutamente
maravilloso.

Reproduction (R)

Todo
iba viento en popa.

Creative shift (CS)

How to Annotate Creativity? Errors

Chicago Lying-in Hospital

Hospital Chicago Lying-in

Error (E)

R: Hospital de Maternidad de Chicago

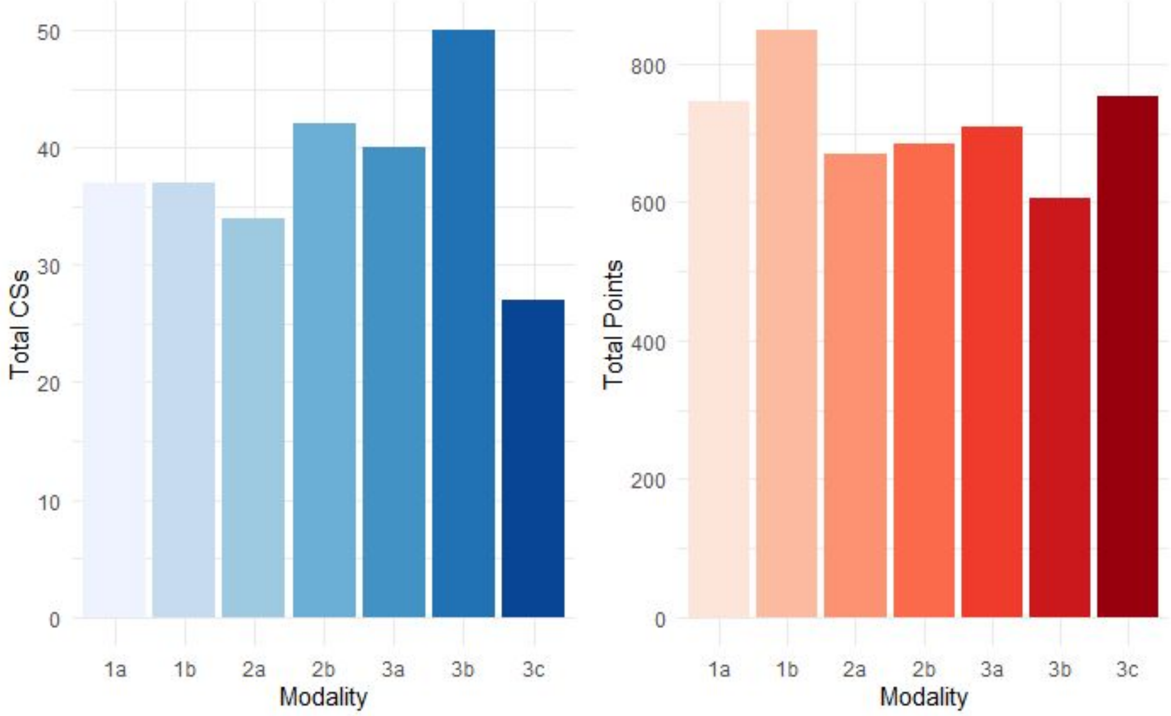
CS: Hospital Materno Infantil de la Ciudad

How to Annotate Creativity? Formula

Guerberof-Arenas and Toral (2022)

$$CI = \left(\underbrace{\frac{\#CSs}{\#UCPs}}_{\text{Novelty}} - \underbrace{\frac{\#error\ points}{\#\ words\ in\ ST}}_{\text{Errors}} \right) * 100$$

Results (Du et al. 2025)

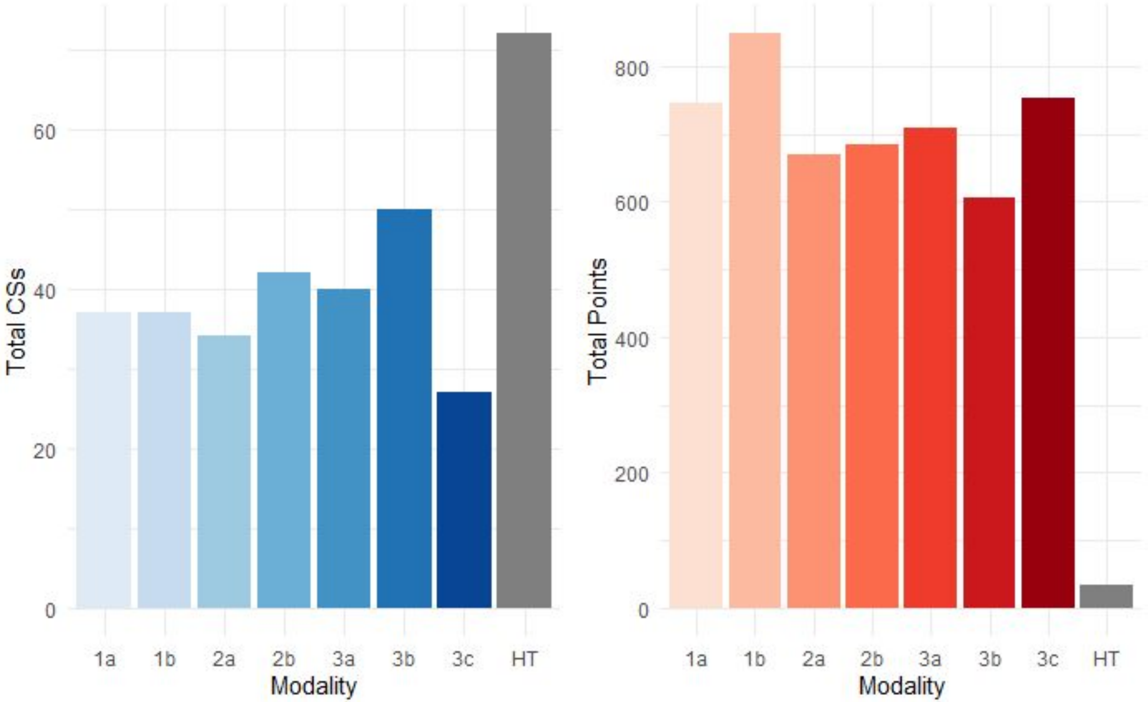


1a-3a: ChatGPT (other)

3b: ChatGPT (optimal)

3c: NMT

Results (Du et al. 2025)



3.2. Natural MT?

“Des études ont montré que parmi les enjeux posés par les outils qui fournissent automatiquement des traductions se trouve la standardisation de la langue qui est générée. On parle ainsi de «machine translationese »”

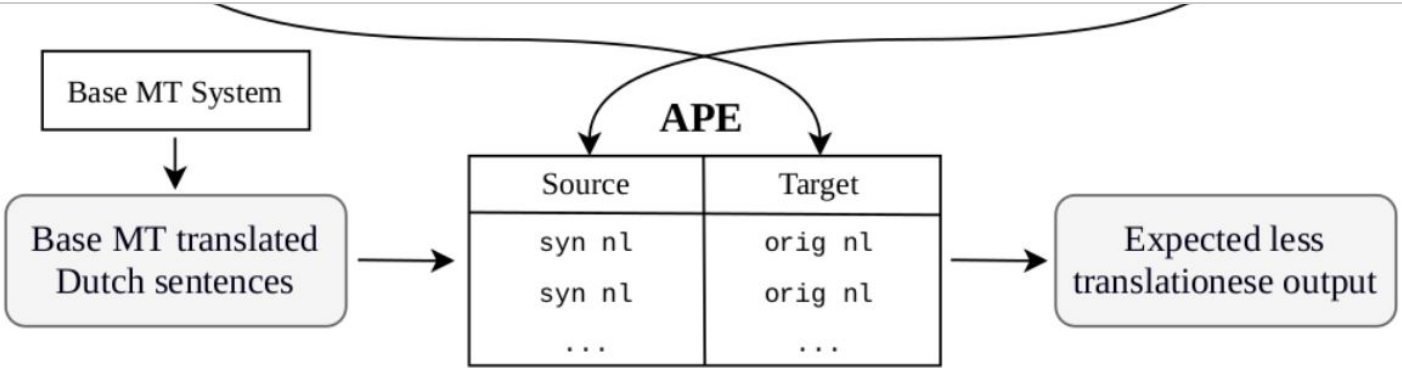
Machine Translationese

- Vanmassenhove et al. (2019). Less lexical variety in MT than HT
    - MT overgenerates frequent words and undergenerates infrequent words
- Webster et al. (2020). Syntactic structure of MT similar to that of ST
- Toral (2019). MT>PE>HT in terms of simplicity and interference from ST
- Loock (2020). Supporting qualitative analyses

Reducing Machine Translationese

- Automatic Post-editing (APE) (Freitag et al., 2019)
- Tagging (Freitag et al., 2022)
- LLMs (Raunak et al., 2023; Li et al. 2025)
- Reinforcement learning (Lai et al. 2025)

APE (Freitag et al. 2019)



Base MT System

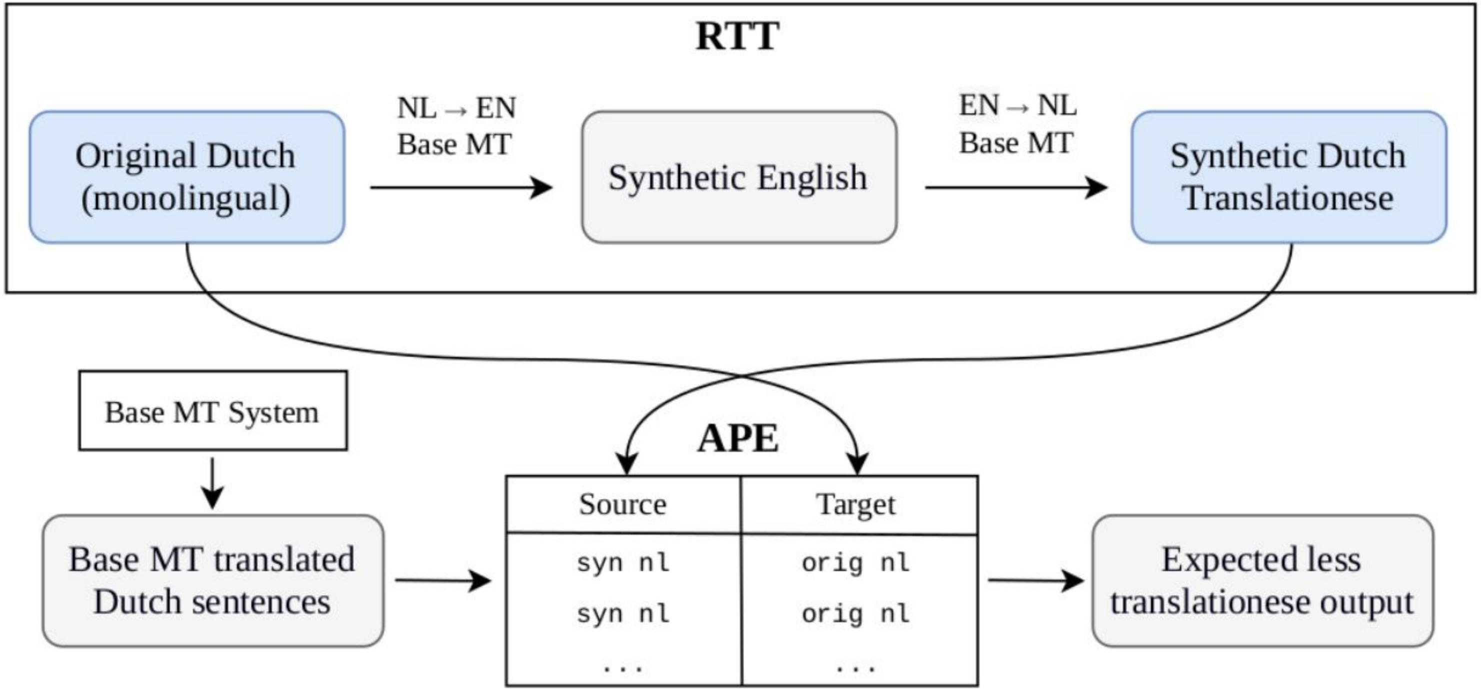
Base MT translated Dutch sentences

APE

Source	Target
`syn nl`	`orig nl`
`syn nl`	`orig nl`
...	...

Expected less translationese output

APE (Freitag et al. 2019)



RTT

Original Dutch (monolingual)

NL → EN Base MT

Synthetic English

EN → NL Base MT

Synthetic Dutch Translationese

Base MT System

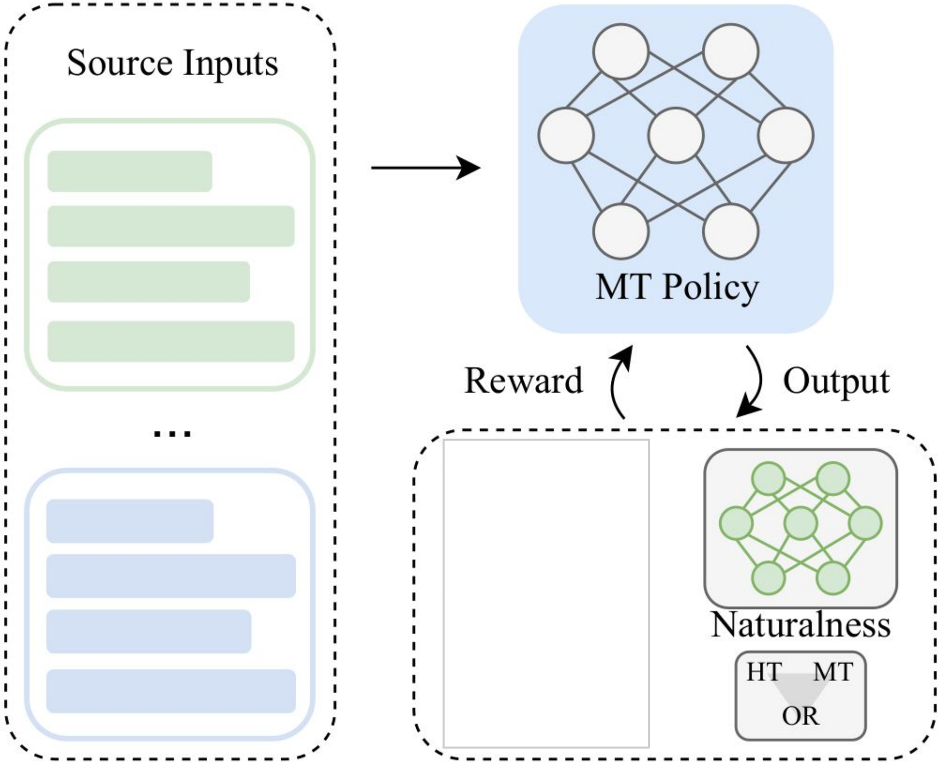
Base MT translated Dutch sentences

APE

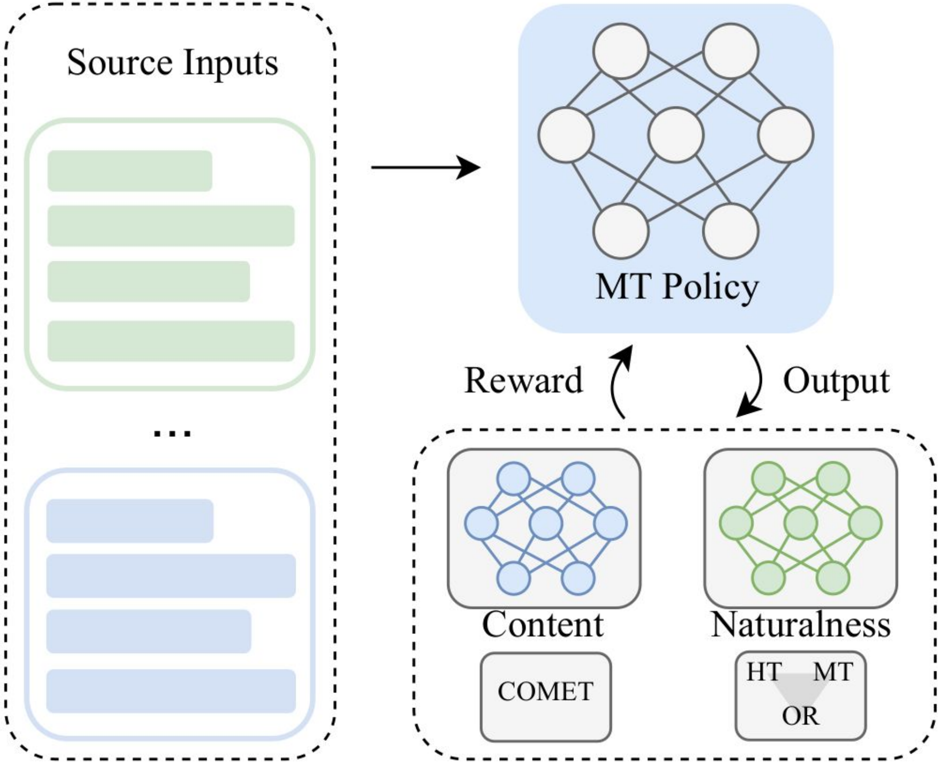
Source	Target
`syn nl`	`orig nl`
`syn nl`	`orig nl`
...	...

Expected less translationese output

Reinforcement Learning (Lai et al. 2025)



Reinforcement Learning (Lai et al. 2025)



Reinforcement Learning (Lai et al. 2025)

	MetricX↓	**HT-OR**	**MT-HT**	**MTLD**
HT				
Base MT				
Tagging				
RW (HT-OR)				
RW (MT-HT)				

Reinforcement Learning (Lai et al. 2025)

	MetricX↓	**HT-OR**	**MT-HT**	**MTLD**
HT	-	32.9	69.3	96.0
Base MT	**2.66**	28.1	18.9	90.4
Tagging	2.87	**33.0**	**42.6**	**95.8**
RW (HT-OR)	2.83	**34.0**	25.5	91.0
RW (MT-HT)	**2.63**	26.1	**26.6**	**93.3**

4. LLMs on Your Machine

“Enfin la question de la créativité en traduction pose la question de l’acte créatif en lien avec la propriété intellectuelle”

LLMs Online. Privacy

When we use a service such as Claude, or ChatGPT…

- … our input **data** travels to a Data Center
- … a gen AI model is queried with our data
- … we get the **result** back

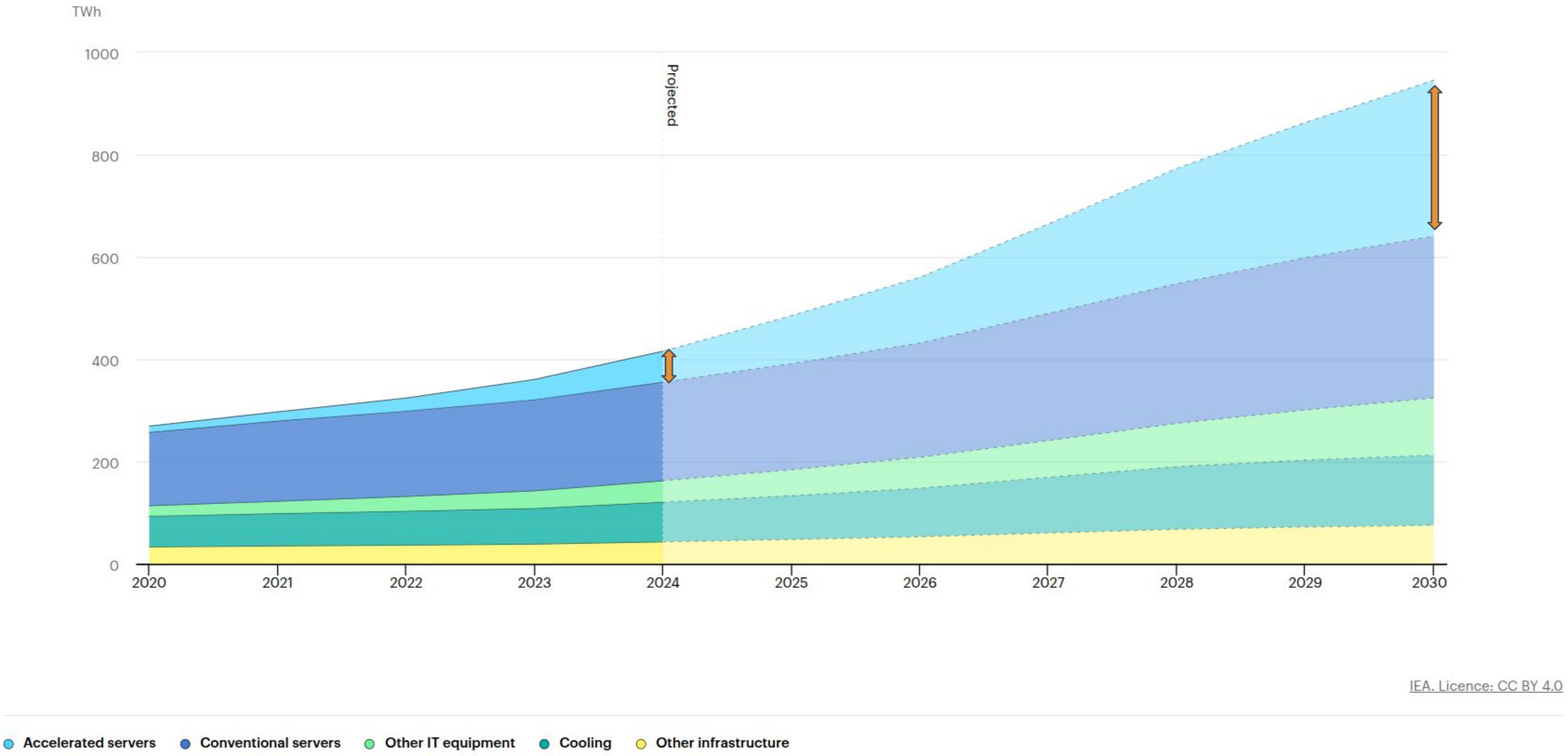
Data can travel safely through the Internet, but…
… **what do** AI providers do with **our data**?

When you don’t pay for a product, it may be that you are the product

LLMs Online. Data centers

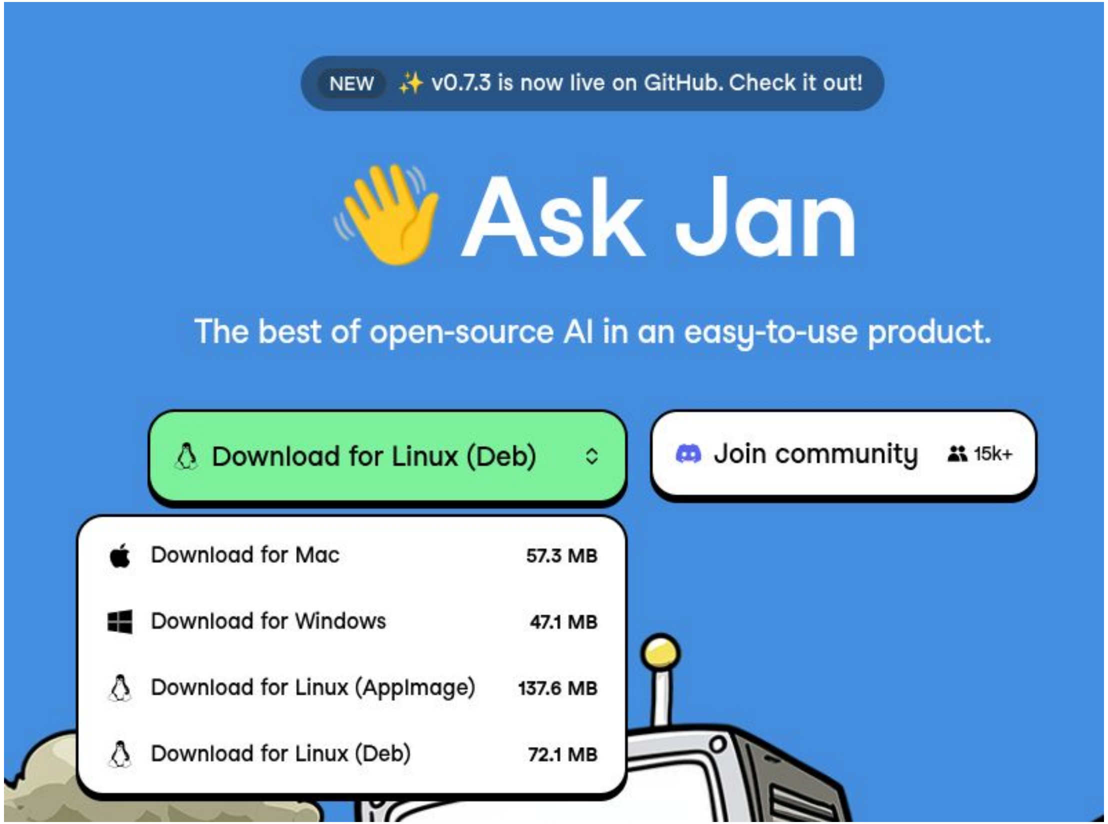
According to *International Energy Agency* (IEA) they consume 3% of the world's energy

LLMs Online. Data centers



IEA. Licence: CC BY 4.0

jan.ai

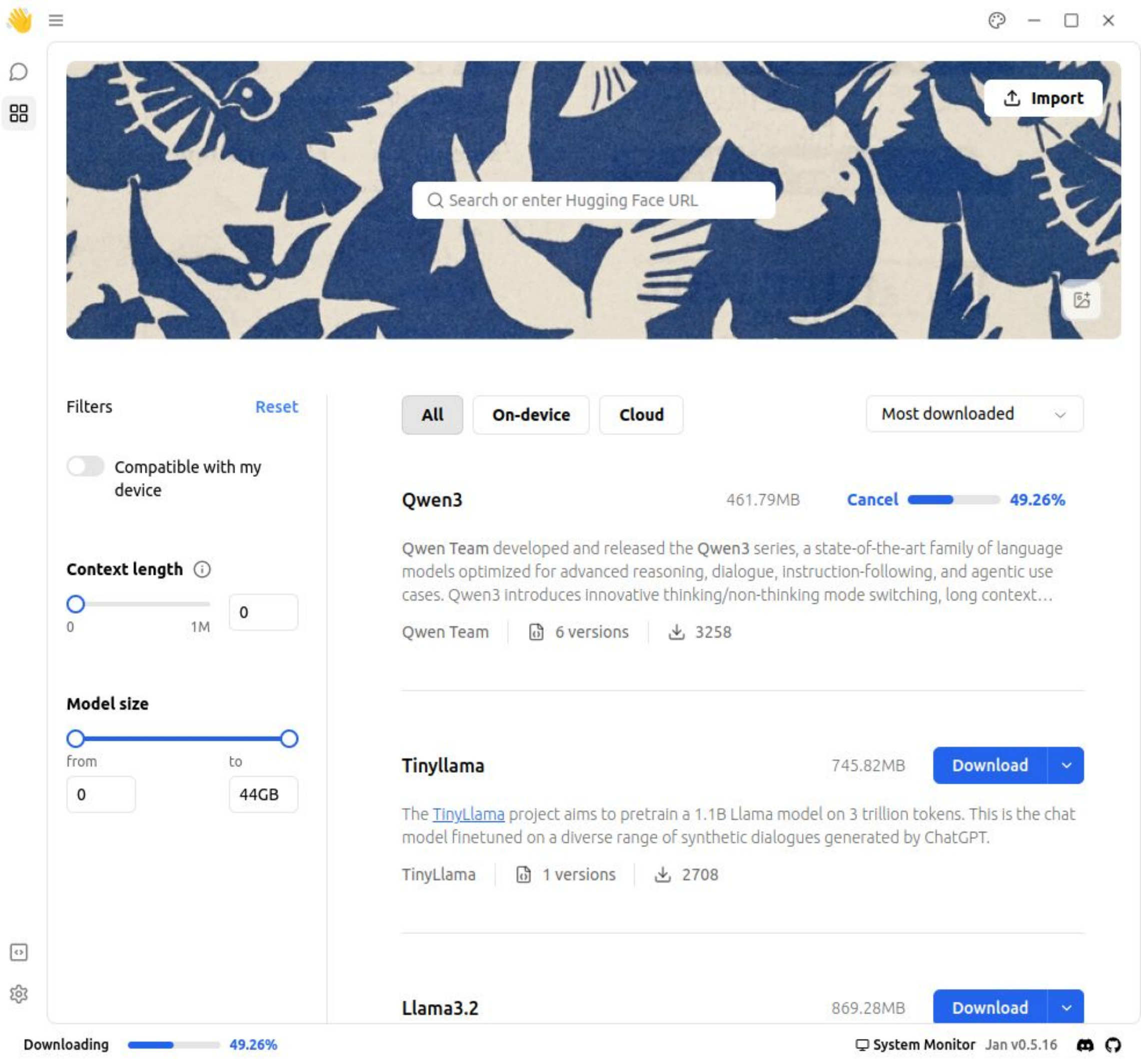


The best of open-source AI in an easy-to-use product.

Download for Mac	57.3 MB
Download for Windows	47.1 MB
Download for Linux (AppImage)	137.6 MB
Download for Linux (Deb)	72.1 MB

Jan. Models

Hugging Face



Qwen3 461.79MB

Qwen Team developed and released the Qwen3 series, a state-of-the-art family of language models optimized for advanced reasoning, dialogue, instruction-following, and agentic use cases. Qwen3 introduces innovative thinking/non-thinking mode switching, long context…

Qwen Team | 6 versions | 3258

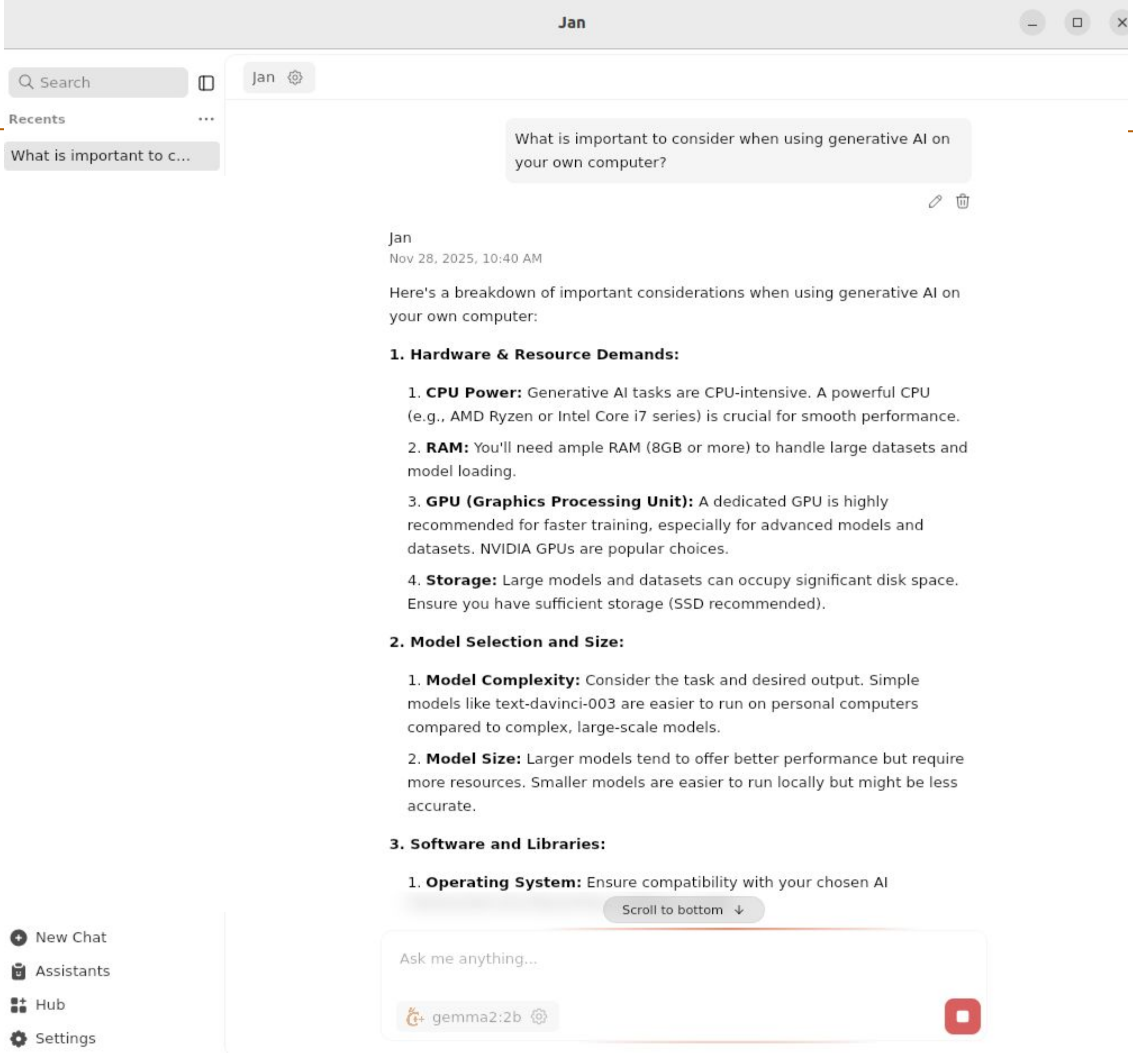
Tinyllama 745.82MB

The TinyLlama project aims to pretrain a 1.1B Llama model on 3 trillion tokens. This is the chat model finetuned on a diverse range of synthetic dialogues generated by ChatGPT.

TinyLlama | 1 versions | 2708

Llama3.2 869.28MB

Jan. Interacting



What is important to consider when using generative AI on your own computer?

Jan
Nov 28, 2025, 10:40 AM

Here's a breakdown of important considerations when using generative AI on your own computer:

1. Hardware & Resource Demands:

1. **CPU Power:** Generative AI tasks are CPU-intensive. A powerful CPU (e.g., AMD Ryzen or Intel Core i7 series) is crucial for smooth performance.
2. **RAM:** You'll need ample RAM (8GB or more) to handle large datasets and model loading.
3. **GPU (Graphics Processing Unit):** A dedicated GPU is highly recommended for faster training, especially for advanced models and datasets. NVIDIA GPUs are popular choices.
4. **Storage:** Large models and datasets can occupy significant disk space. Ensure you have sufficient storage (SSD recommended).

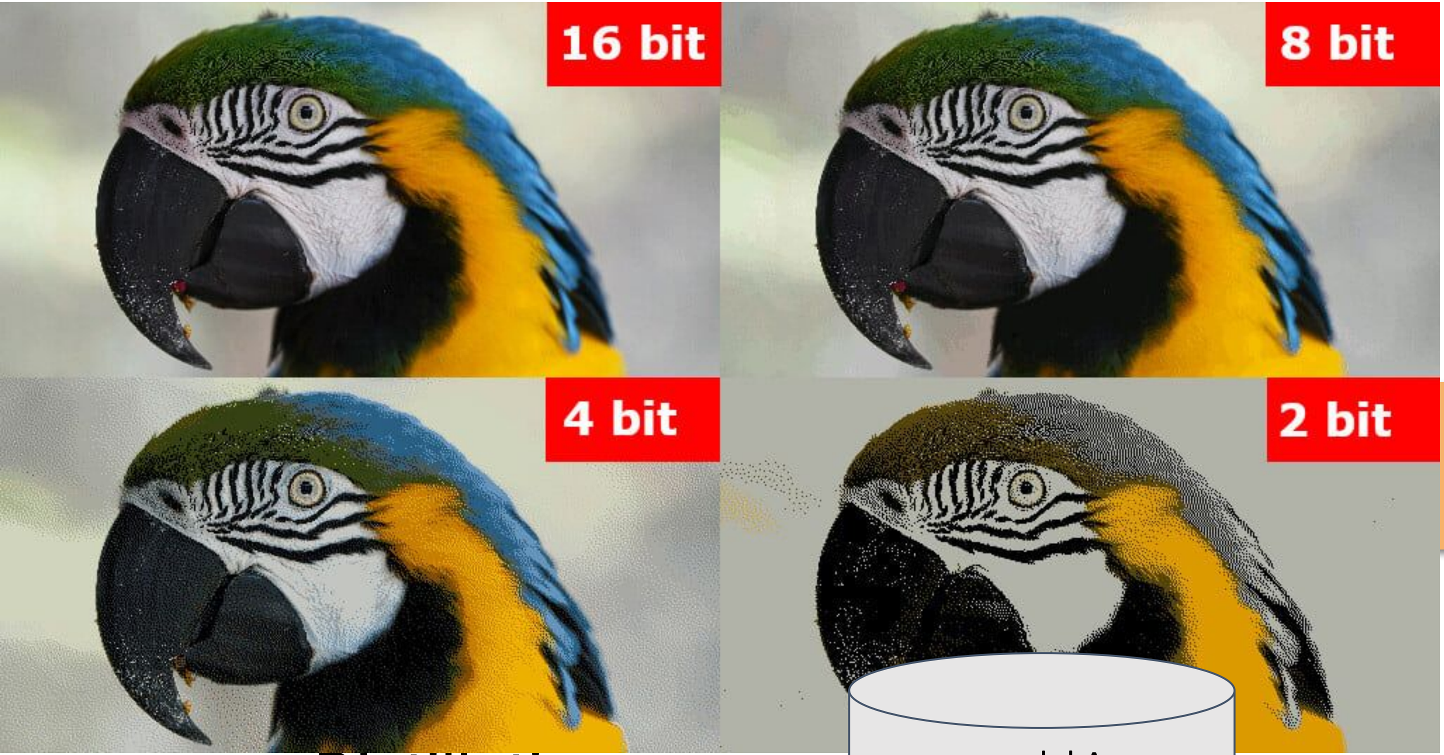
2. Model Selection and Size:

1. **Model Complexity:** Consider the task and desired output. Simple models like text-davinci-003 are easier to run on personal computers compared to complex, large-scale models.
2. **Model Size:** Larger models tend to offer better performance but require more resources. Smaller models are easier to run locally but might be less accurate.

3. Software and Libraries:

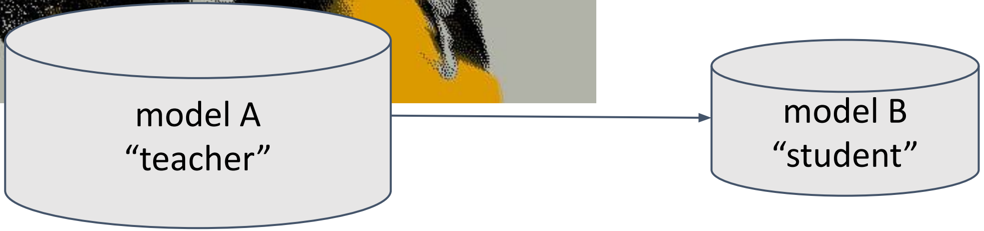
1. **Operating System:** Ensure compatibility with your chosen AI

LLMs Offline. Efficiency



parameter

- Distillation



model A
“teacher”

model B
“student”

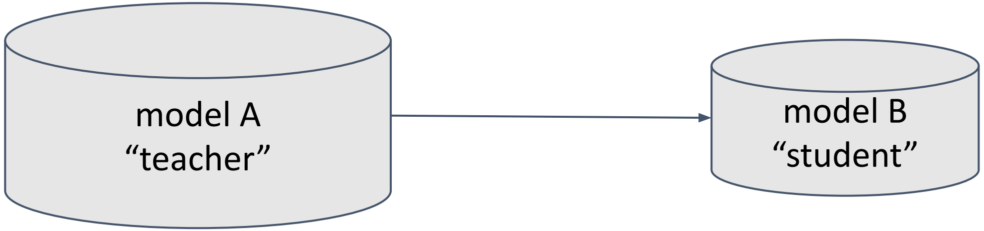
LLMs Offline. Efficiency

2 techniques to have more efficient models

- Quantization

Use less bits for each parameter
E.g. FP32 → INT8

- Distillation



Which LLM can I run on my machine?

Generation speed (tokens/s) of Qwen3 models (Toral et al. 2026)

Quantization: 8 bits (1.7B and 8B), 4 bits (32B and 235B)

Computer	**RAM**	**GPU RAM**	**1.7B**	**8B**	**32B**	**235B-A22B**
2014 laptop	8 GB		4			
2019 laptop	16 GB		14	4		
2023 laptop	96 GB	?	105	31	13	1
2018 server	128 GB	12 GB	157	55	3	
2018 server	128 GB	24 GB	153	55	22	2

Adapting an Offline LLM

How can we adapt a model with our own data?

- Retrieval Augmented Generation (RAG)
- Supervised Fine Tuning (SFT)

More on this

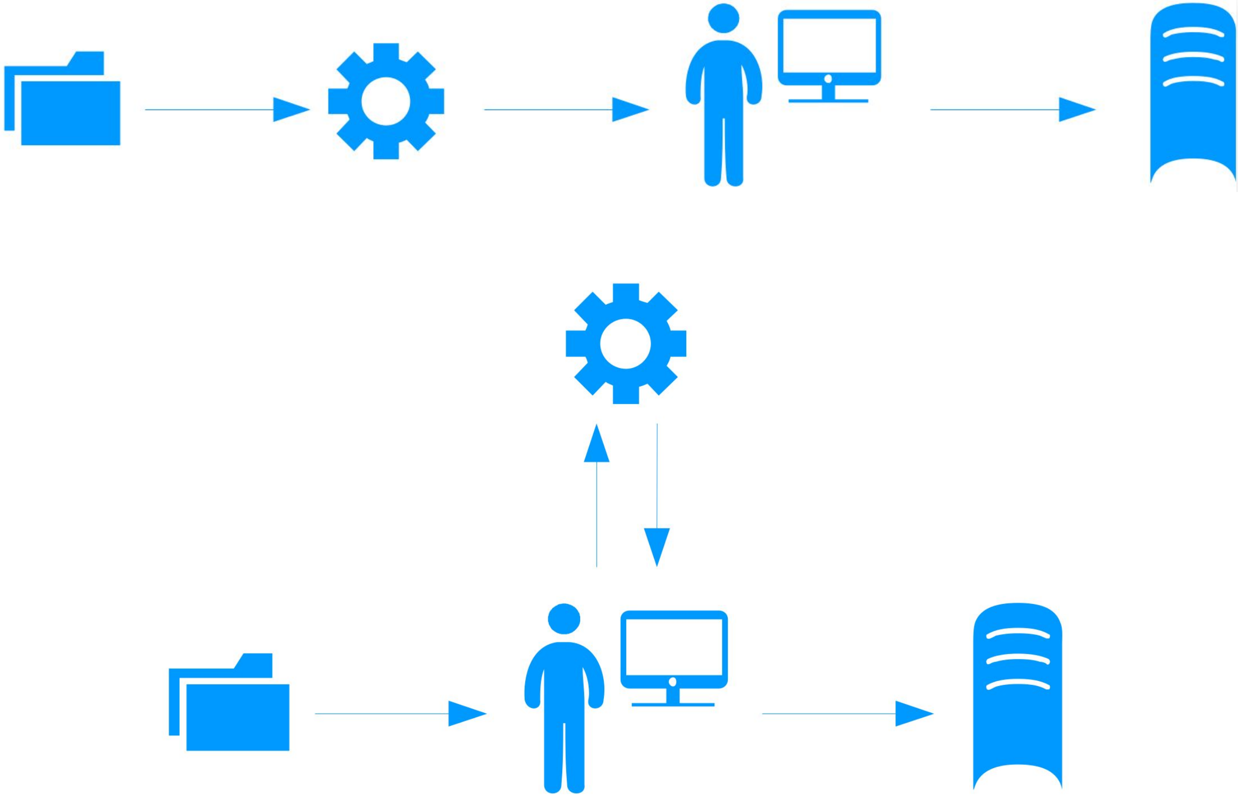
Upcoming LT-LiDER book chapter

- A. Toral, A. van Cranenburgh, M. Esplà, A. Guerberof-Arenas. How can translators use and customise their own private LLM?

Other related chapters

- Key Concepts in LLMs
- How can translators and students make the most out of LLMs?
- Python for translators

“Comment rester créatif face à la machine ?”



Take home

- LLMs have clear advantages over NMT… but there are also disadvantages
- Translations by LLMs more creative than by NMT, but far from professional translators
- Offline optimised LLMs can be competitive for translation

Merci!

https://antoniotor.al/tq26.pdf

GenAI for Translators: An Introduction to LLMs and their Usage

Antonio Toral

LT-LiDER

UA Universitat d'Alacant
Universidad de Alicante

References

- Bayer-Hohenwarter. ‘Translational Creativity: How to Measure the Unmeasurable’. In Behind the Mind: Methods, Models and Results in Translation Process Research. 2009.
- Bayer-Hohenwarter. ‘Creative Shifts as a Means of Measuring and Promoting Translational Creativity’. Meta 56 (3): 663–92. 2011.
- Du et al. Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish. 2025
- Freitag, Caswell, and Roy. Ape at scale and its implications on mt evaluation biases. arXiv 2019.
- Freitag, Vilar, Grangier, Cherry, and Foster. A natural diet: Towards improving naturalness of machine translation output. Findings of ACL. 2022.
- Guerberof-Arenas, Ana, and Antonio Toral. Creativity in translation: Machine translation as a constraint for literary texts. Translation Spaces, v. 11 n. 2. 2022.

References

- Kocmi et al. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. WMT 2025.
- Lai et al. Multi-perspective Alignment for Increasing Naturalness in Neural Machine Translation. ACL 2025.
- Li et al. Lost in Literalism: How Supervised Training Shapes Translationese in LLMs. ACL 2025.
- Loock. No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment. The Journal of specialised translation (JoSTrans) 34. 2020.
- Pangeanic. Which is better for my use case (neural) NMT or LLM translation? Our White Book. Blog post 2025.
- Raunak et al. Do gpts produce less literal translations? ACL 2023.
- Runco and Jaeger. The standard definition of creativity. Creativity Research Journal, 24(1), 92–96. 2012
- Toral. Post-editese: an exacerbated translationese. MT Summit 2019.

References

- Toral, van Cranenburgh, Esplà, Guerberof-Arenas. How can translators use and customise their own private LLM? 2026. *Under Review.*
- Vanmassenhove, Shterionov, and Gwilliam. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. EACL 2021
- Wang et al. Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding. WMT 2025.
- Webster et al. Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. Informatics. Vol. 7. No. 3. 2020.
- Zheng et al. 2025. Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task. WMT

Additional Slides

Some Conclusions from WMT 2025

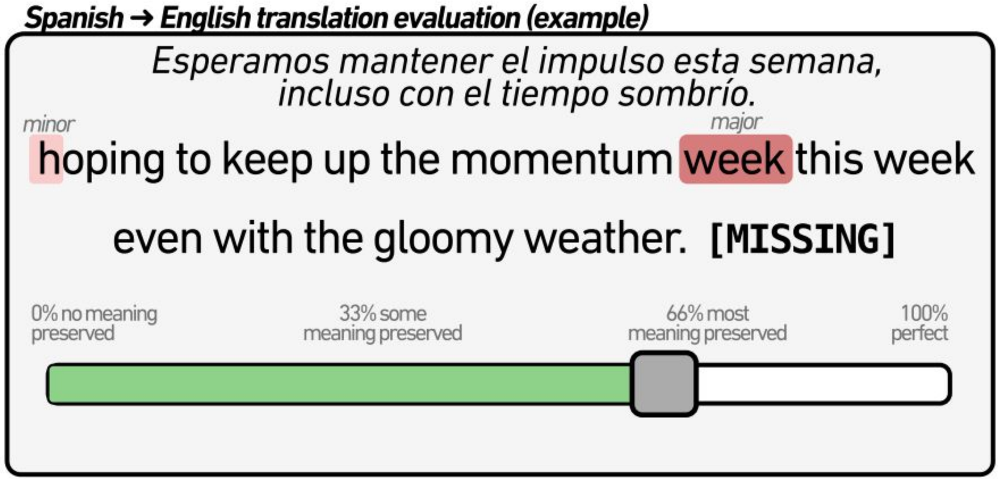
Translation

- Automatic metrics are biased
- HT is not always the best evaluated: only for 6 out of 15 language pairs

Automatic evaluation

- Good results with general-purpose LLMs and detailed prompting
- Poor results from trained neural metrics (e.g. COMET)

Human Evaluation at WMT 2025



Spanish → English translation evaluation (example)

Esperamos mantener el impulso esta semana, incluso con el tiempo sombrío.

hoping to keep up the momentum week this week even with the gloomy weather. `[MISSING]`

Figure 1: Stylized annotation user interface with Error Span Annotation (ESA). The annotator first marks errors with minor and major severity and then assigns a final score. This is more robust than asking for score directly.[1]

Translationese

The language used in translated texts tends to have different characteristics from the language of original texts (Baker, 1995; Toury, 1995; Teich, 2003):

- Explanations
- Normalizations
- Simplifications
- Interference

Not necessarily problematic, but MT exacerbates them!

Reinforcement Learning (Lai et al. 2025)

$$\mathcal{L}(\theta;\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\beta\mathcal{L}_{nl} + \mathcal{L}_{rw}]$$

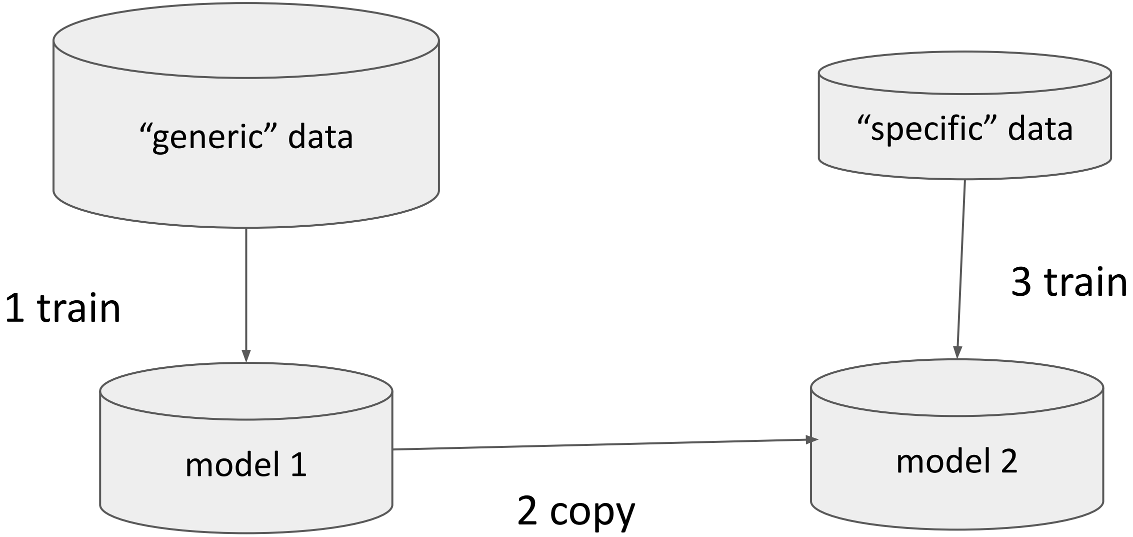
$$\mathcal{L}_{rw} = -\frac{1}{m}\sum_{i=1}^{m} r(\hat{y})\ldots$$

$$r(\hat{y}) = \begin{cases} 0 & \text{if } r_t = 0 \text{ or } r_c = 0 \\ \frac{2}{1/r_t + 1/r_c} & \text{otherwise} \end{cases}$$

$$r_c(\hat{y}) = \begin{cases} 0 & \text{if } \mathrm{C}(x,y,\hat{y}) < \sigma_c \\ \mathrm{C}(x,y,\hat{y}) & \text{otherwise} \end{cases}$$

$$r_t(\hat{y}) = \begin{cases} 0 & \text{if } p(t_1|\hat{y};\phi) < \sigma_t \\ p(t_1|\hat{y};\phi) & \text{otherwise} \end{cases}$$

Supervised Fine Tuning (SFT)



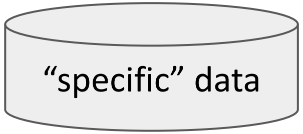
Retrieval Augmented Generation (RAG)

Prompt

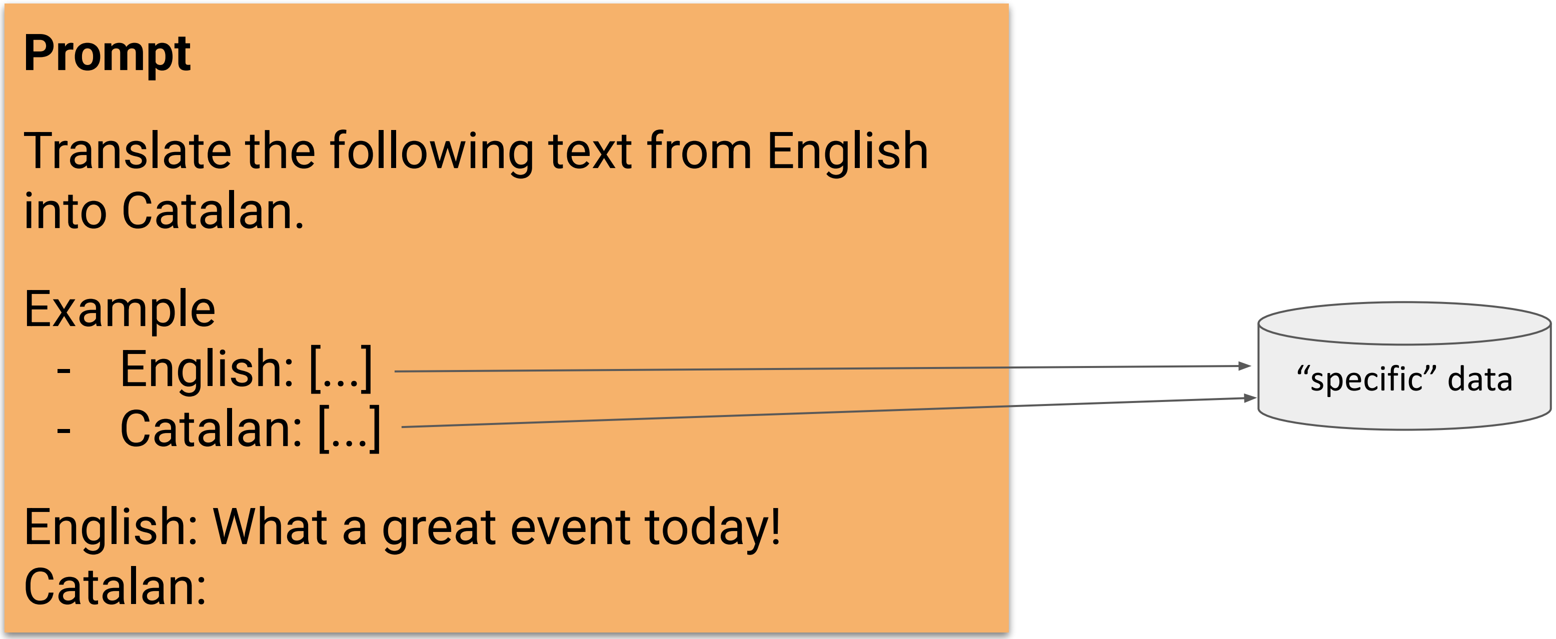
Translate the following text from English into Catalan.

Example
- English: [...]
- Catalan: [...]

English: What a great event today!
Catalan:



Retrieval Augmented Generation (RAG)



Prompt

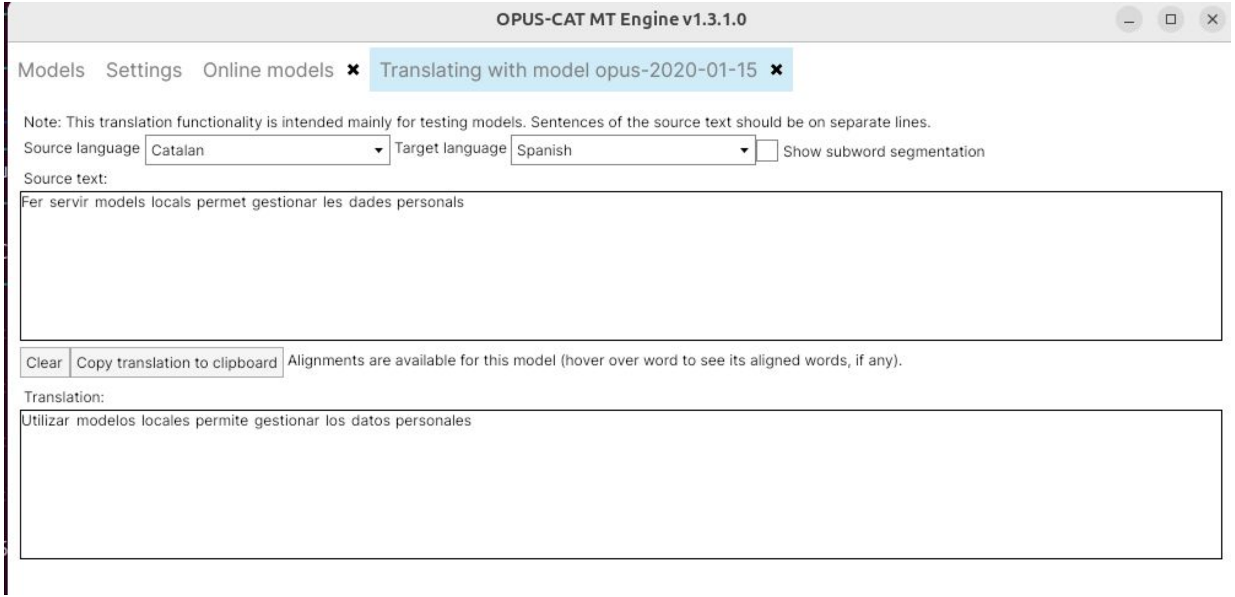
Translate the following text from English into Catalan.

Example
- English: [...]
- Catalan: [...]

English: What a great event today!
Catalan:

“specific” data

NOT ONLY GENAI



OPUS-CAT MT Engine v1.3.1.0
Models
Settings
Online models
Translating with model opus-2020-01-15
Clear
Copy translation to clipboard

Note: This translation functionality is intended mainly for testing models. Sentences of the source text should be on separate lines.

Source language: Catalan

Target language: Spanish

☐ Show subword segmentation

Source text:

Fer servir models locals permet gestionar les dades personals

Alignments are available for this model (hover over word to see its aligned words, if any).

Translation:

Utilizar modelos locales permite gestionar los datos personales