



# Traduction automatique

## Grands modèles de langue autorégressifs

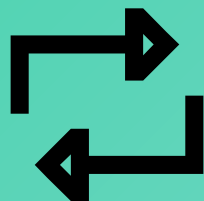
**Enzo Doyen**

enzo.doyen@unistra.fr

2025-07-20

## **Plan**

- I. Principes de base des modèles de langue autorégressifs
- II. Émergence des modèles à base d'instructions
- III. Grands modèles de langue et applicabilité à la traduction



# **I. Principes de base des modèles de langue autorégressifs**

# Modèles de langue

Un modèle de langue est, traditionnellement, un modèle statistique qui prédit la probabilité d'une séquence de mots.

Soit  $V$  l'ensemble du vocabulaire, un modèle de langue  $p$  est une fonction qui associe à chaque séquence de mots  $w_1, \dots, w_n \in V$  une probabilité (un nombre entre 0 et 1) :

$$p(w_1, \dots, w_n)$$

# Modèles de langue : exemple

Prenons un vocabulaire  $V = \{\text{la, verte, maison, verrerie, est}\}$ .

Différentes probabilités pourront être assignées par le modèle de langue :

- $p(\text{la, maison, est, verte}) = 0.08$
- $p(\text{la, verrerie, est, verte}) = 0.01$
- $p(\text{verrierie, est, maison, la}) = 0.0001$

Les probabilités sont apprises à partir d'entraînement sur de grands corpus.

# Modèles de langue autorégressifs

Un modèle de langue autorégressif est un modèle de langue qui prédit la probabilité d'un mot en fonction des mots précédents.

En appliquant la règle de la dérivation en chaîne (*chain rule*), on peut écrire la probabilité d'une séquence de mots comme suit :

$$= P(w_1 \dots w_n) = P(w_1)P(w_2 \mid w_1) \dots P(w_n \mid w_1, \dots, w_{n-1})$$

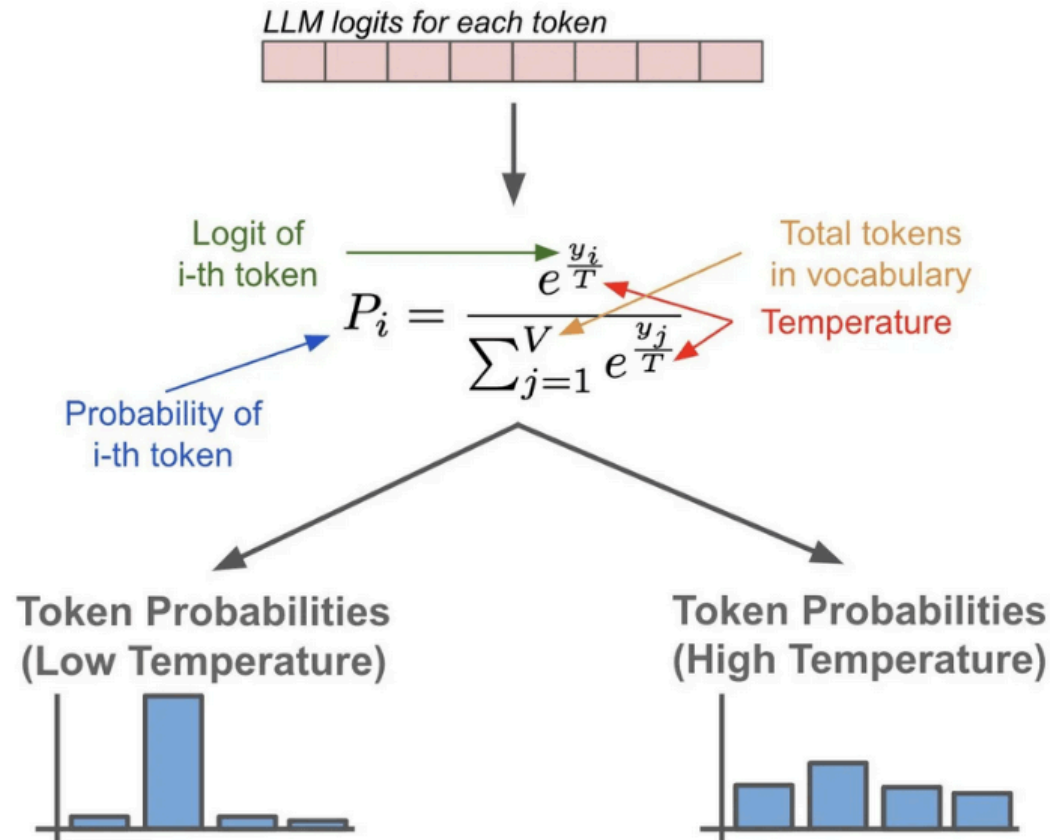
De manière simplifiée :

$$= \prod_{t=1}^n P(w_t \mid \mathbf{w}_{<t})$$

# Température

La **température**  $T$  est un paramètre utilisé pour **contrôler la diversité des tokens générés** par le modèle de langue ; plus la température est élevée, plus le modèle est susceptible de générer des tokens moins probables par rapport à ce qui a été observé dans les données d'entraînement, ce qui ajoute de la **variation**.

# Softmax with Temperature



Source : [https://medium.com/@amansinghalml\\_33304/temperature-llms-b41d75870510](https://medium.com/@amansinghalml_33304/temperature-llms-b41d75870510)



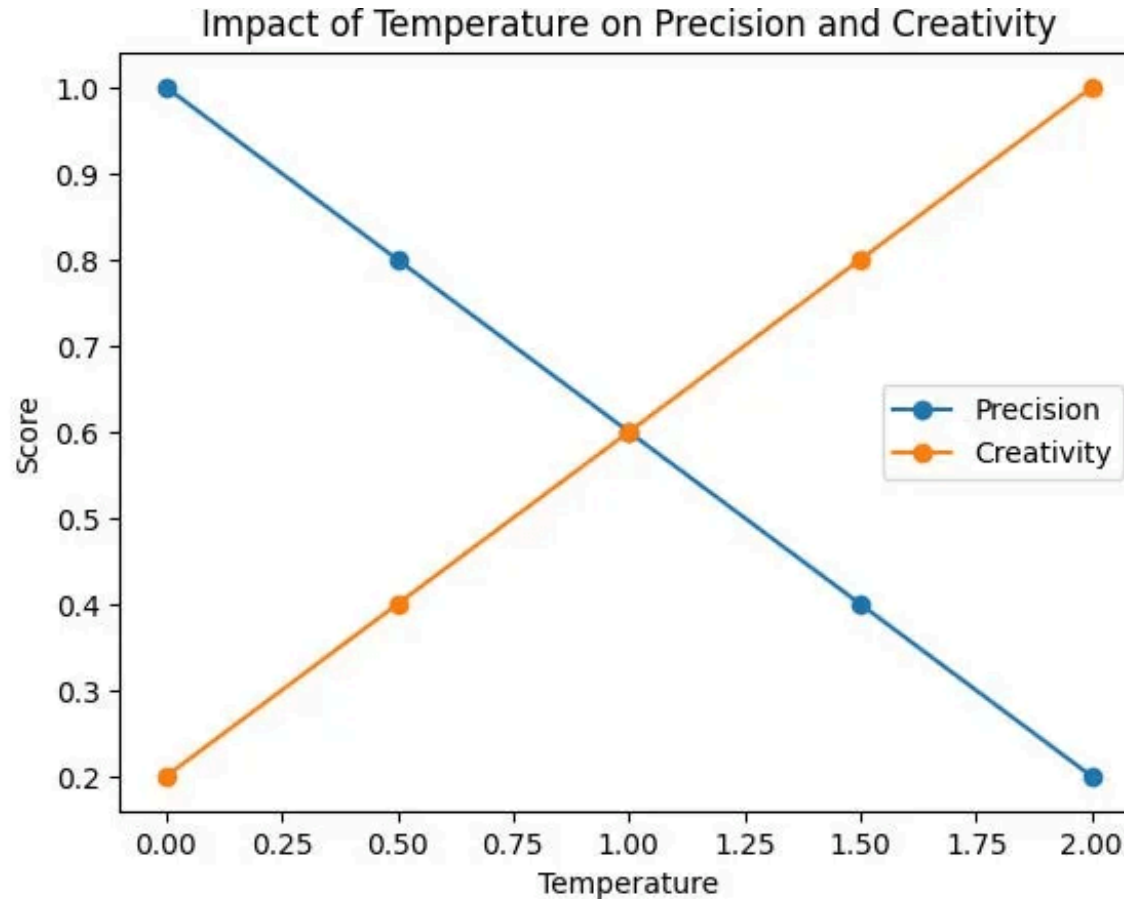
# Température

La **température**  $T$  est un paramètre utilisé pour **contrôler la diversité des tokens générés** par le modèle de langue ; plus la température est élevée, plus le modèle est susceptible de générer des tokens moins probables par rapport à ce qui a été observé dans les données d'entraînement, ce qui ajoute de la **variation**.

$T = 0$  : choisit toujours le token le plus probable.

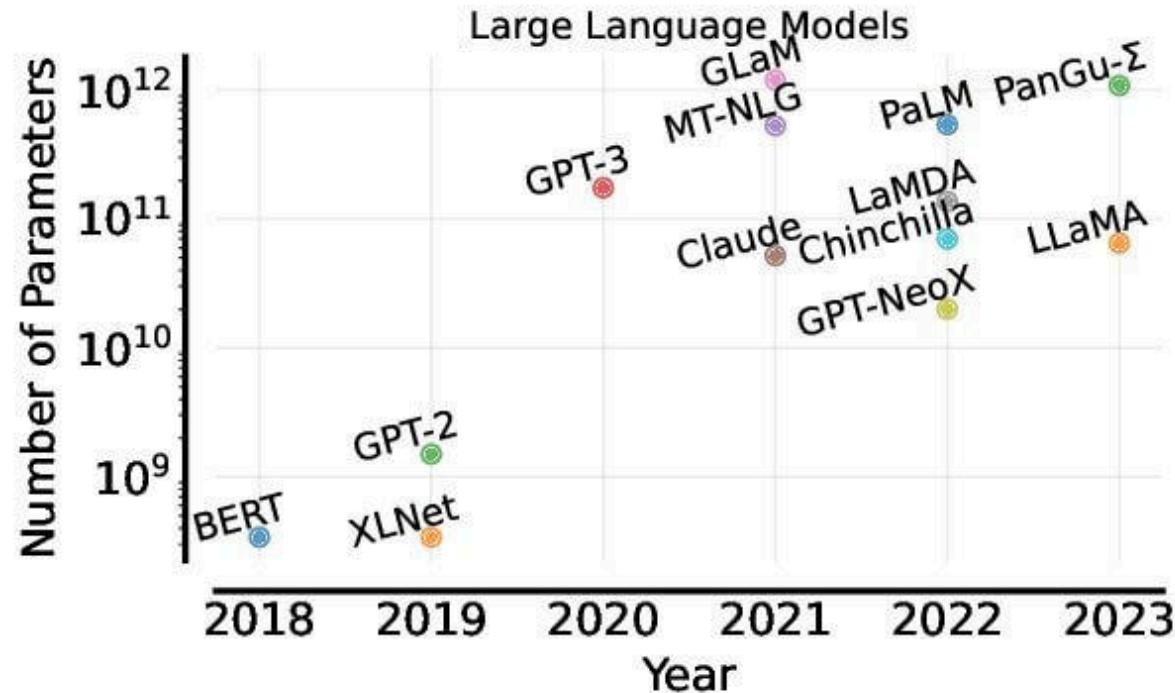
$T = 1$  : distribution de probabilité normale.

$T > 1$  : distribution de probabilité plus uniforme, donc plus de diversité.



Source : <https://nihar-palem.medium.com/understanding-temperature-in-language-models-llms-67079f1d6193>

# Taille des modèles de langue et évolution



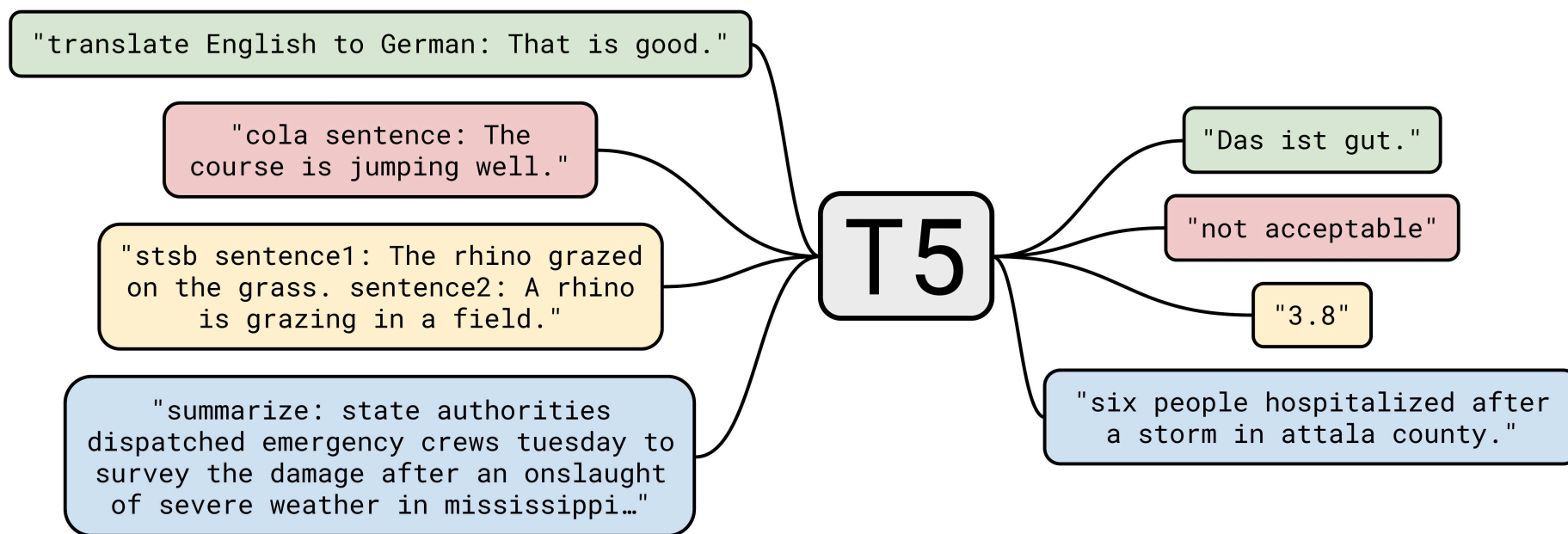
Source : **Agarwal et al. (2023)**



## **II. Émergence des modèles à base d'instructions**

# Modèles de langue multitâches

À partir de 2019, développement de modèles de langue Transformer séquence-à-séquence (comme T5), capables de répondre à plusieurs types de tâches différentes.

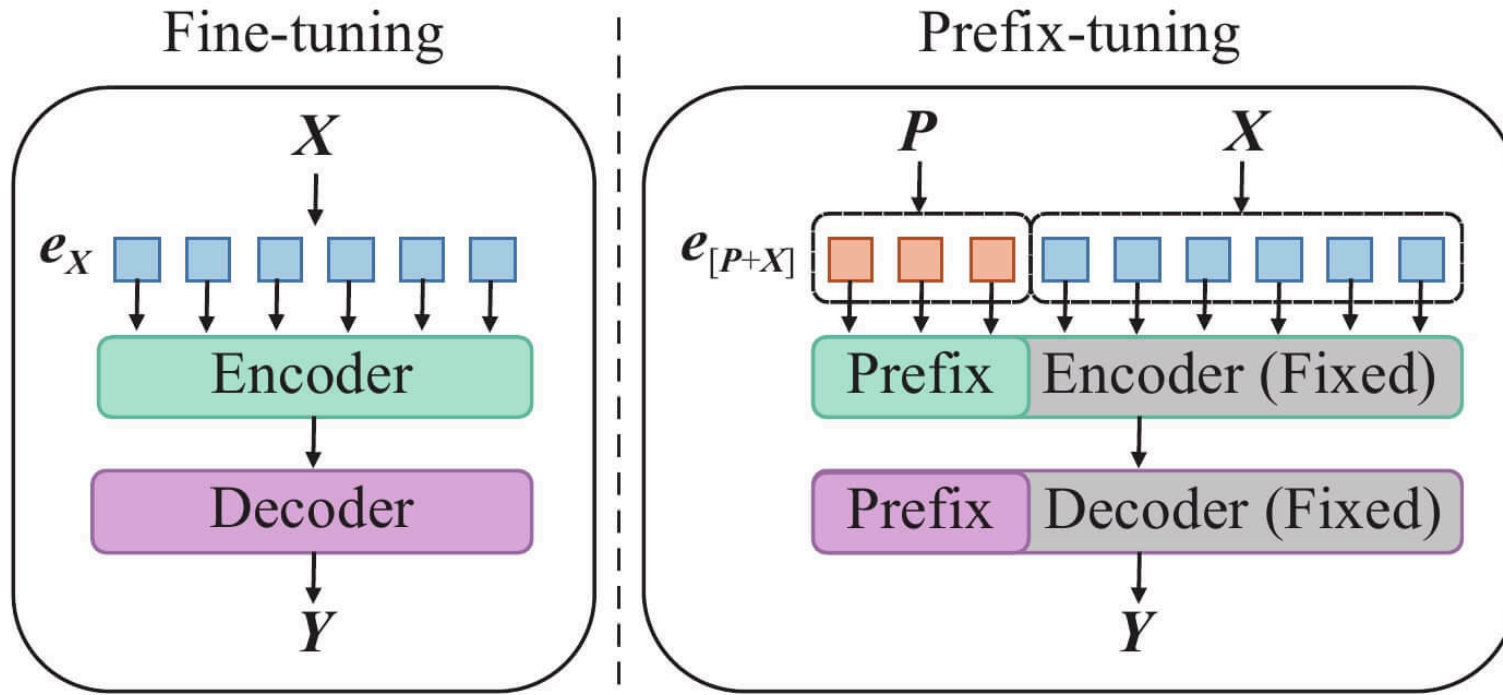


Source : Raffel et al. (2020)

# Modèles de langue multitâches

- ♦ Émergence du concept d'« instruction » (ou « prompt ») pour guider le modèle de langue dans la tâche à accomplir, sous le nom de **prefix-tuning** (Li et Liang, 2021).
- ♦ Tokens « virtuels » qui conditionnent les réponses du modèle : ce sont en fait des vecteurs appris pendant l'entraînement et utilisés comme paires clé/valeur dans le mécanisme d'attention.
- ♦ Les poids du modèle restent inchangés (contrairement à de l'affinage).

# Prefix-tuning



Source : **Chen et al. (2023)**



# Few-shot prompting (Brown et al., 2020)

## Zero-Shot

No Examples

Prompt:

```
Translate the following  
English text to French:  
"The weather is beautiful  
today."
```

```
Le temps est magnifique  
aujourd'hui.
```

## One-Shot

1 Example

Prompt:

```
Example:  
English: "Hello, how are  
you?"  
French: "Bonjour, comment  
allez-vous ?"
```

```
Translate the following  
English text to French:  
"The weather is beautiful  
today."
```

```
Le temps est magnifique  
aujourd'hui.
```

## Few-Shot

3+ Examples

Prompt:

```
Examples:  
English: "Hello, how are  
you?"  
French: "Bonjour, comment  
allez-vous ?"
```

```
English: "I love reading  
books."  
French: "J'adore lire des  
livres."
```

```
English: "The cat is  
sleeping on the sofa."  
French: "Le chat dort sur  
le canapé."
```

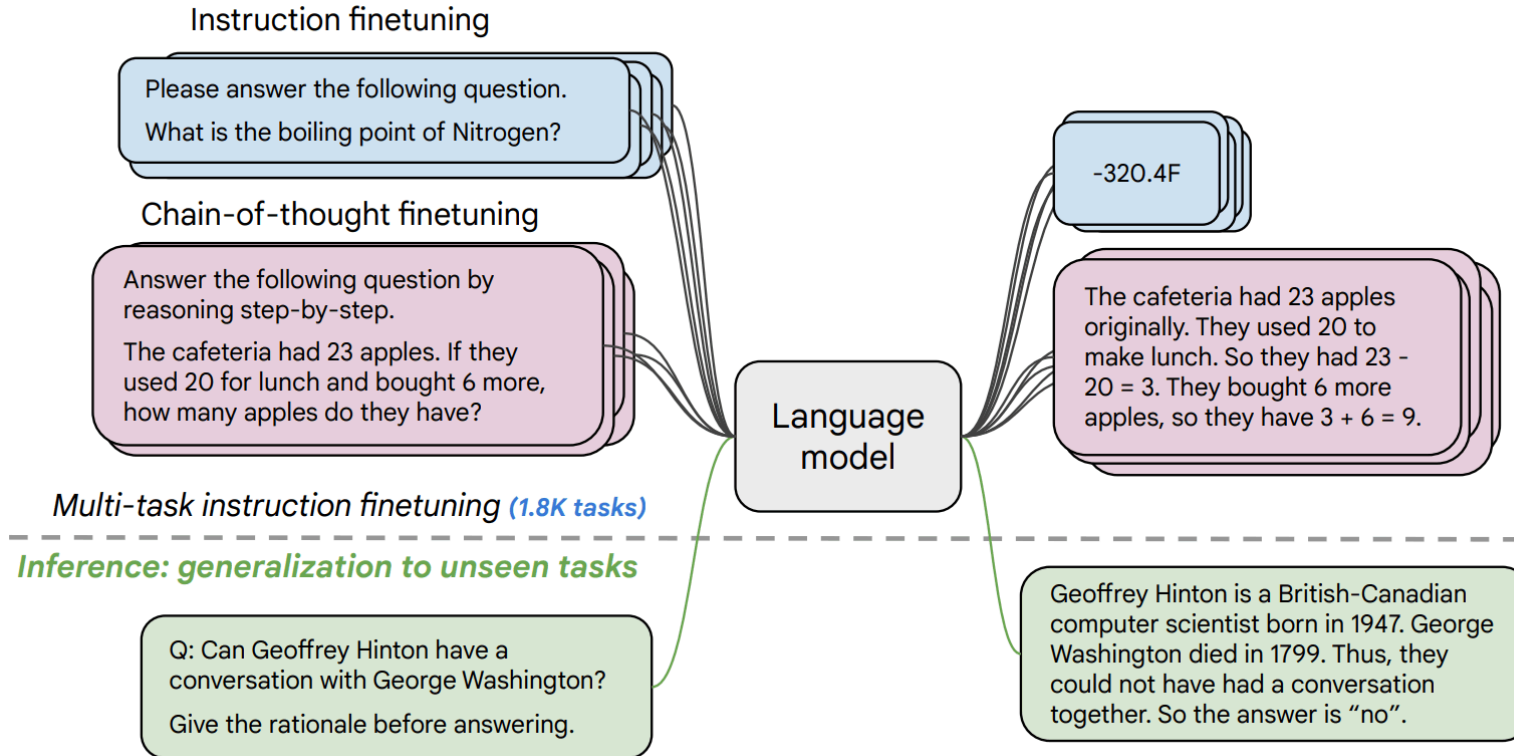
```
Translate the following  
English text to French:  
"The weather is beautiful  
today."
```

```
Le temps est magnifique  
aujourd'hui.
```

# Modèles à base d'instructions

- ♦ Modèles de langue préentraînés, puis affinés sur des tâches spécifiques en utilisant de grands jeux de données de questions-réponses.

# Modèles à base d'instructions





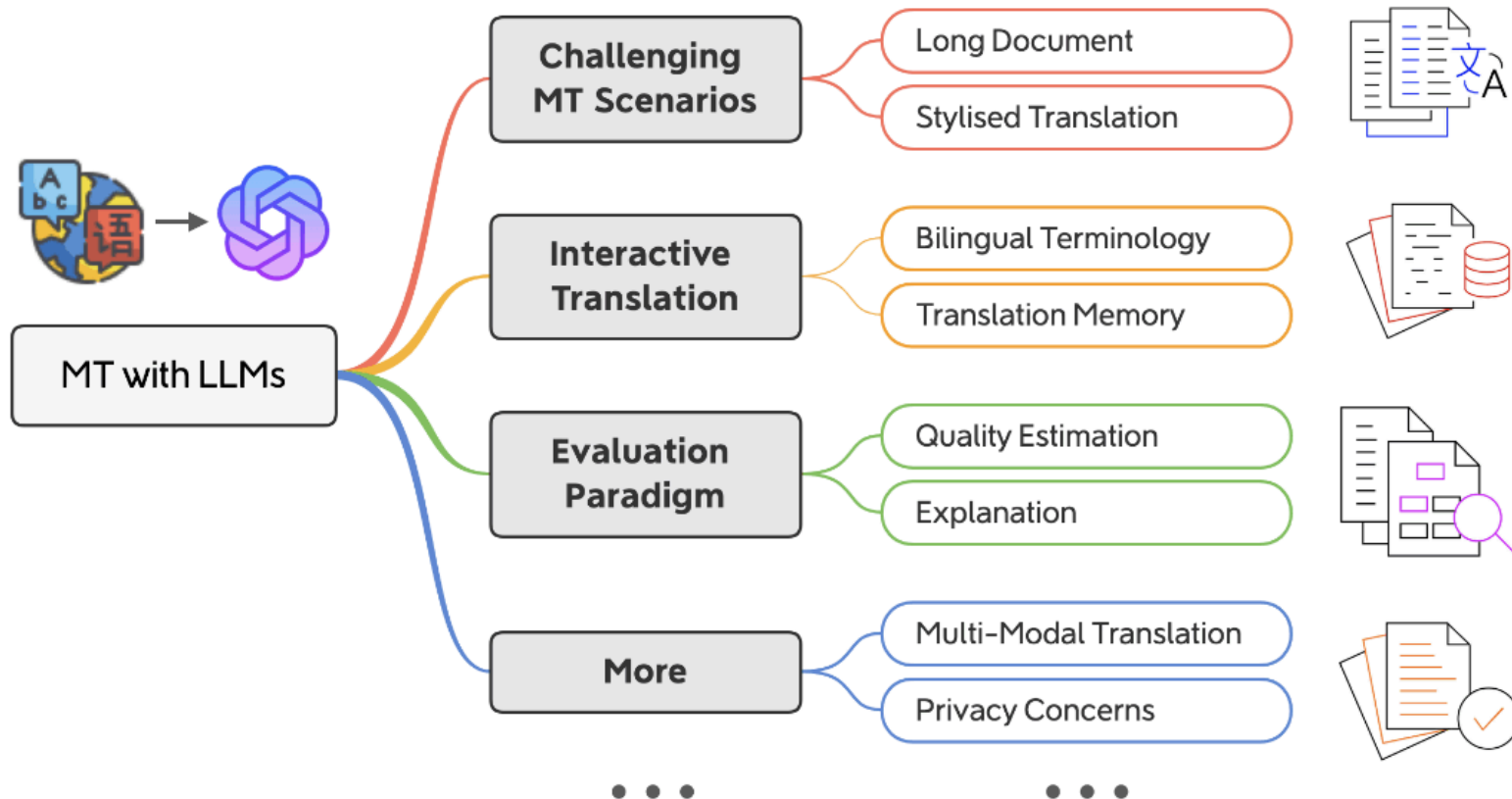
### **III. Grands modèles de langue et applicabilité à la traduction**

# **Grands modèles de langue : la panacée ?**

## **A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models**

**Chenyang Lyu<sup>1</sup>, Zefeng Du<sup>2</sup>, Jitao Xu<sup>3</sup>, Yitao Duan<sup>3</sup>, Minghao Wu<sup>4</sup>,  
Teresa Lynn<sup>1</sup>, Alham Fikri Aji<sup>1</sup>, Derek F. Wong<sup>2</sup>, Longyue Wang<sup>5</sup>**

Source : **Lyu et al. (2024)**

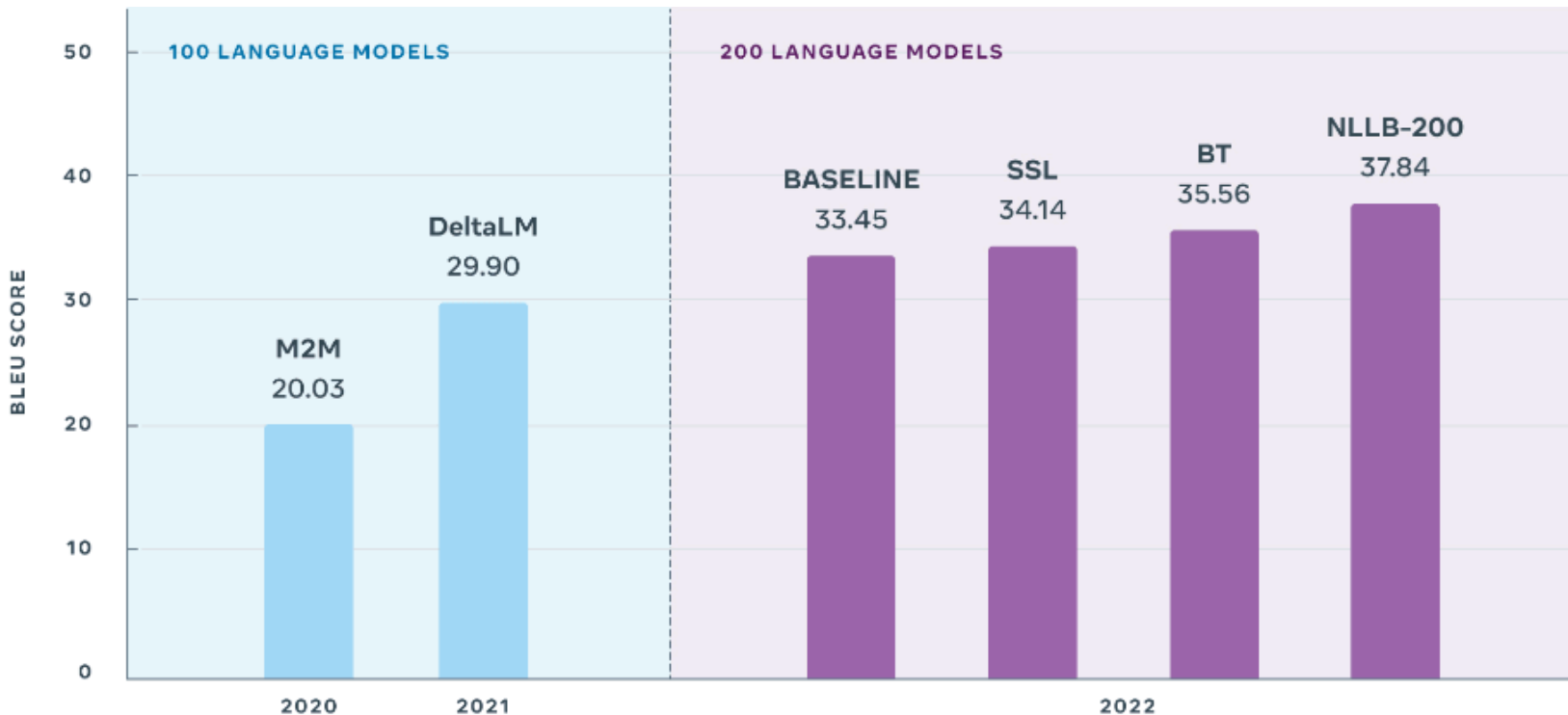


Source : **Lyu et al. (2024)**

# **NLLB-200 (No Language Left Behind) de Meta : modèle de langue de traduction avec 200 langues**

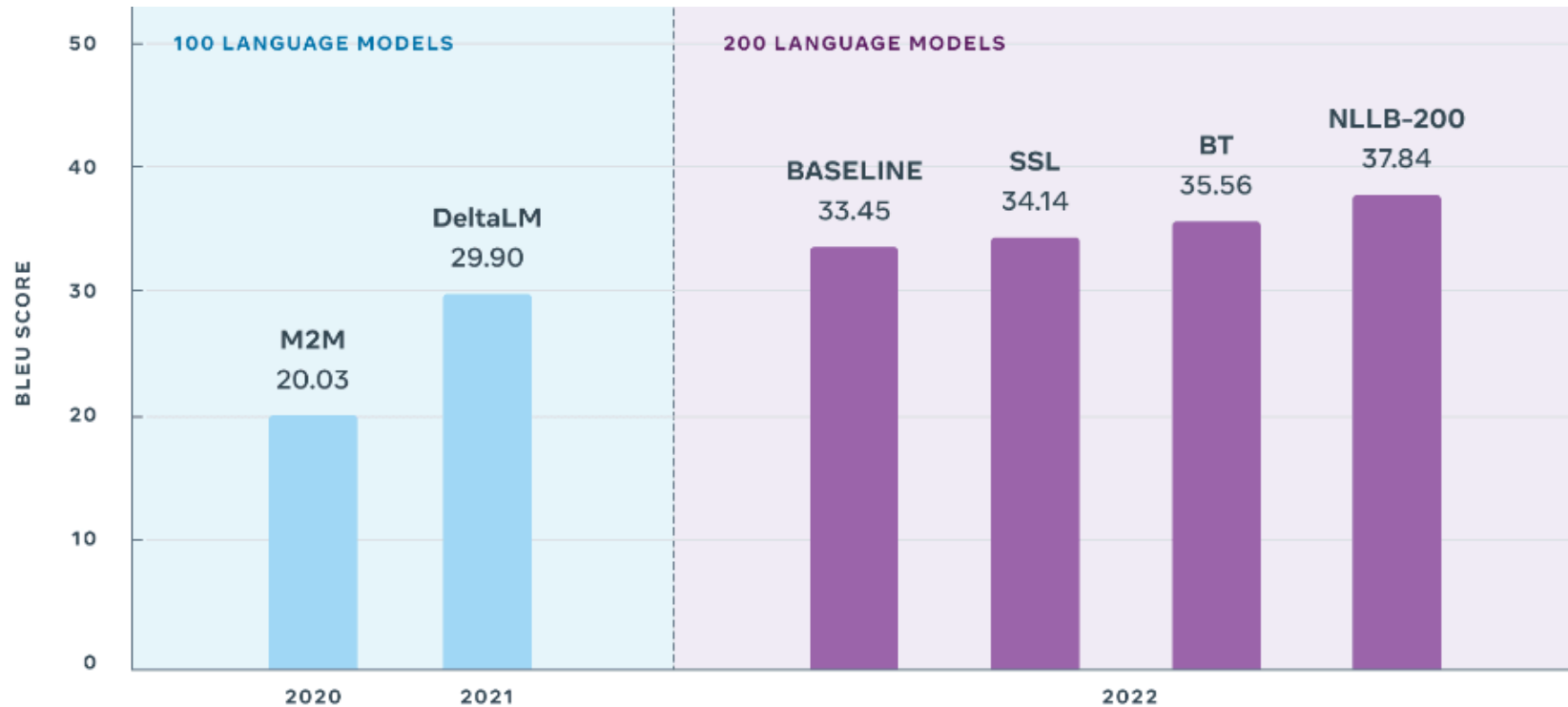
Research

200 languages within a single AI model: A  
breakthrough in high-quality machine  
translation



This graphic shows average BLEU score on FLORES-101 translations to and from English into 100 languages. On the left there are two published state-of-the-art models, M2M and Delta LM, that support 100 languages. Models on the right support 200 languages: A baseline





This graphic shows average BLEU score on FLORES-101 translations to and from English into 100 languages. On the left there are two published state-of-the-art models, M2M and Delta LM, that support 100 languages. Models on the right support 200 languages: A baseline

# 🧑 Implémentation *human-in-the-loop*




Example:




# 🧑‍💻 Implémentation *human-in-the-loop* : intégration aux plateformes de traduction

Sélectionnez le sujet principal de votre événement ou de votre annonce. La catégorie que vous sélectionnez doit généralement correspondre à celle que vous décrivez dans le titre de votre événement et dans l'illustration de votre image de couverture.

 Amazon Translate

Sélectionnez le sujet principal de votre événement ou annonce. La catégorie que vous sélectionnez devrait généralement correspondre à ce que vous décrivez dans le titre de votre événement et dans l'œuvre pour votre image de couverture.

 Crowdin Translate (beta)

Sélectionnez le thème principal de votre événement ou annonce. La catégorie choisie doit généralement correspondre à la description du titre de votre événement et à celle de votre image de couverture.

 Google AutoML Translate - test 1

## ⚡ Technique : *role prompting*

Act as a translator with 20+ years of experience  
in translating from French to English...

You are an expert in translating texts in the field  
of computer science...

# ⚡ Technique : *few-shot prompting* (Brown et al., 2020)

## Zero-Shot No Examples

Prompt:

```
Translate the following
English text to French:
"The weather is beautiful
today."
```

```
Le temps est magnifique
aujourd'hui.
```

## One-Shot 1 Example

Prompt:

```
Example:
English: "Hello, how are
you?"
French: "Bonjour, comment
allez-vous ?"
```

```
Translate the following
English text to French:
"The weather is beautiful
today."
```

```
Le temps est magnifique
aujourd'hui.
```

## Few-Shot 3+ Examples

Prompt:

```
Examples:
English: "Hello, how are
you?"
French: "Bonjour, comment
allez-vous ?"
```

```
English: "I love reading
books."
French: "J'adore lire des
livres."
```

```
English: "The cat is
sleeping on the sofa."
French: "Le chat dort sur
le canapé."
```

```
Translate the following
English text to French:
"The weather is beautiful
today."
```

```
Le temps est magnifique
aujourd'hui.
```

# ⚡ Technique : traduction itérative (Chen et al., 2024)

| Mode                              | Prompt   |
|-----------------------------------|--|
| <i>Translate</i>                  | Source: <code>\${source}</code><br>Please give me a translation in <code>\${lang}</code> without any explanation.  |
| <i>Refine</i>                     | Source: <code>\${source}</code><br>Translation: <code>\${prev_translation}</code><br>Please give me a better <code>\${lang}</code> translation without any explanation.  |
| <i>Refine</i> <sub>Contrast</sub> | Source: <code>\${source}</code><br><b>Bad</b> translation: <code>\${prev_translation}</code><br>Please give me a better <code>\${lang}</code> translation without any explanation.   |
| <i>Refine</i> <sub>Random</sub>   | Source: <code>\${source}</code><br><b>Bad</b> translation: <code>\${random_target}</code> if first-round, else <code>\${prev_translation}</code><br>Please give me a better <code>\${lang}</code> translation without any explanation. |

# ⚡ Technique : traduction itérative (Feng et al., 2025)

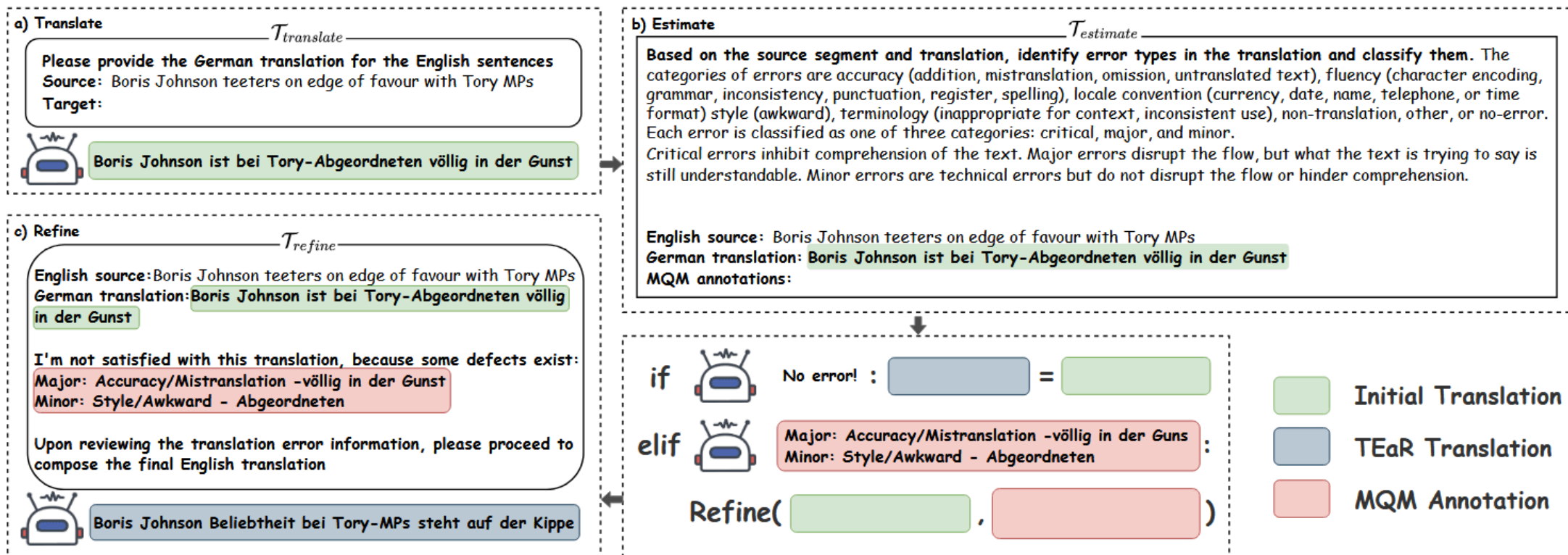
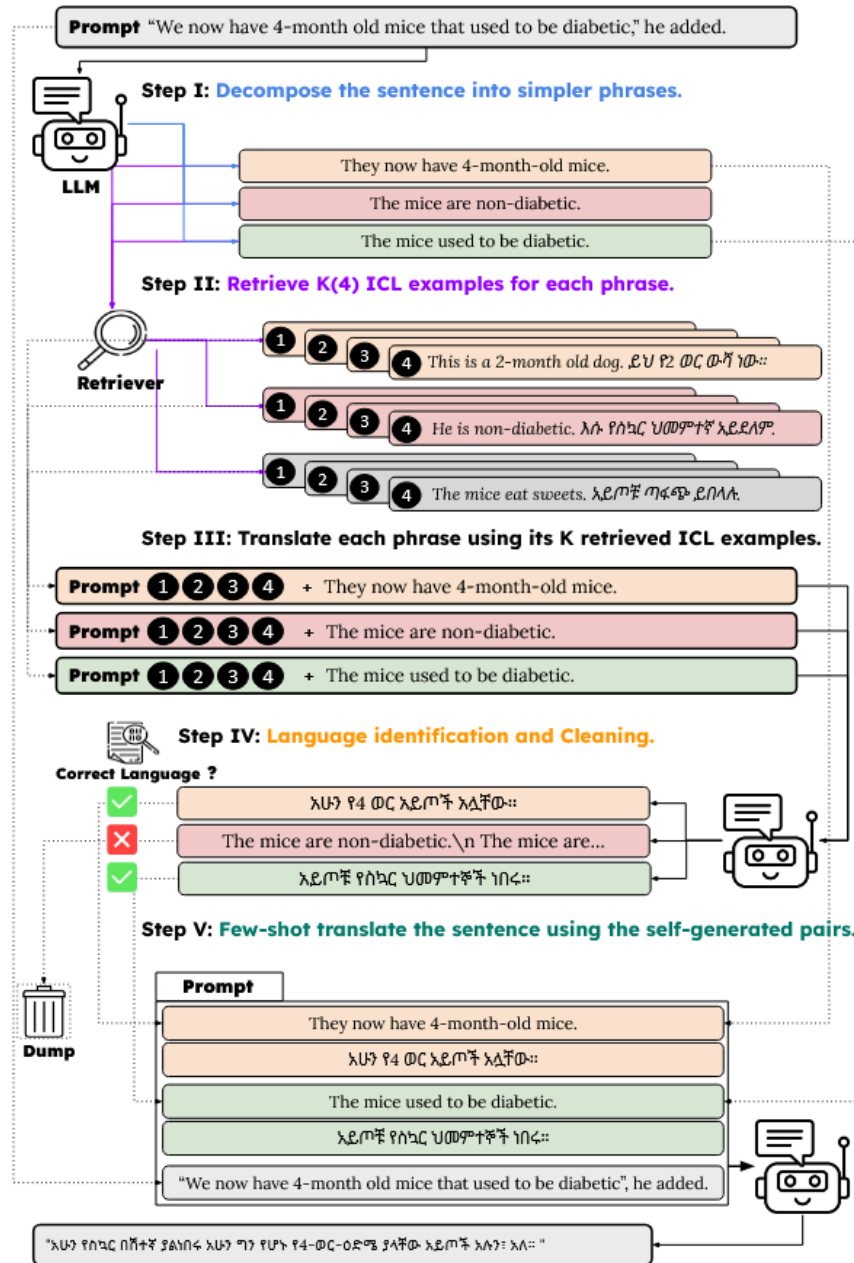


Figure 2: TEaR framework involves three steps: *Translate*, *Estimate*, and *Refine*.

# ⚡ Technique : traduction compositionnelle pour langues peu dotées (zebaze et al., 2025)





# ⚡ Technique : raisonnement (Feng et al., 2025)

|           | EN-JA   | EN-ZH  | DE-ZH  |
|-----------|---|--|--|
| Step 0    | <p>User: A few more sensors in general would be nice I guess.</p> <p>Assistant: <b>&lt;think&gt; I will translate the given English sentence into Japanese. The sentence is "A few more sensors in general would be nice I guess." I will break down the sentence into its main components and find the appropriate Japanese words and phrases for each part. &lt;/think&gt;</b><br/><b>&lt;translate&gt; 一般に、少し更多的センサーがあると良いと思います。&lt;/translate&gt;</b></p> | <p>User: #PilotTraining #Flying #TeaganAir #PA28 #Pilot #FlightSchool</p> <p>Assistant: <b>&lt;think&gt; I will translate the given text from English to Chinese. The text contains hashtags and mentions of a pilot training program, flying, a company called Teagan Air, a specific aircraft model (PA28), and a pilot. I will translate each part of the text accordingly. &lt;/think&gt;</b><br/><b>&lt;translate&gt; #飞行员培训 #飞行 #泰根航空 #PA28 #飞行员 #飞行学校&lt;/translate&gt;</b></p> | <p>User: Heutzutage schreiben Menschen Nachrichten auf Computerbildschirmen, ohne je in die Nähe eines Anspitzers kommen zu müssen.</p> <p>Assistant: <b>&lt;think&gt; To translate the given German sentence into Chinese, I will first break down the sentence into its main components and then find the appropriate Chinese equivalents for each component. The main components are: "Heutzutage" (today), "schreiben" (write), "Menschen" (people), "Nachrichten" (messages), "auf Computerbildschirmen" (on computer screens), "ohne je" (without ever), "in die Nähe" (near), "eines Anspitzers" (a pencil), and "kommen zu müssen" (to have to come). &lt;/think&gt;</b><br/><b>&lt;translate&gt; 现在, 人们可以在电脑屏幕上写信, 而无需接近铅笔。&lt;/translate&gt;</b></p> |
| Step 1600 | <p>User: A few more sensors in general would be nice I guess.</p> <p>Assistant: <b>&lt;think&gt; 私はユーザーから、「User: あ、一般にセンサーが少し増えたらいいなあ」という英語の文を翻訳してほしいと頼まれました。この文は、ユーザーがセンサーの数を増やしてほしいと願っていることを表しています。私はこの文を日本語に翻訳します。「あ、一般にセンサーをもう少し増やしてもいいかなあ」という意味です。&lt;/think&gt;</b><br/><b>&lt;translate&gt; 「一般に、センサーをもう少し増やしてもいいかなあ。」&lt;/translate&gt;</b></p>   | <p>User: #PilotTraining #Flying #TeaganAir #PA28 #Pilot #FlightSchool</p> <p>Assistant: <b>&lt;think&gt; 我在思考如何将这段英文文本“#PilotTraining #Flying #TeaganAir #PA28 #Pilot #FlightSchool”翻译成中文。文本中包含多个标签, 这些标签可能来自社交媒体帖子或文章, 用于描述与飞行员培训、飞行、TeaganAir航空公司、PA28飞机型号、飞行员、飞行学校等相关的内容。翻译时, 我会将每个标签逐一翻译成中文, 保留其原始含义和用途。翻译后的文本内容如下:</b><br/><b>#飞行员培训 #飞行 #泰根航空 #PA28型飞机 #飞行员 #飞行学校&lt;/think&gt;</b><br/><b>&lt;translate&gt; #飞行员培训 #飞行 #泰根航空 #PA28型飞机 #飞行员 #飞行学校&lt;/translate&gt;</b></p>      | <p>User: Heutzutage schreiben Menschen Nachrichten auf Computerbildschirmen, ohne je in die Nähe eines Anspitzers kommen zu müssen.</p> <p>Assistant: <b>&lt;think&gt; 我在思考如何将这段德文文本“Heutzutage schreiben Menschen Nachrichten auf Computerbildschirmen, ohne je in die Nähe eines Anspitzers kommen zu müssen.”翻译成中文。文本中用户引用了一段话, 指出如今人们可以在电脑屏幕上撰写消息, 而无需接近任何针状物。原文指出, 现在人们可以通过电脑屏幕发送消息, 而不需要亲自接触到针状物 (可能是指针式打字机或其他类似设备)。这段文本可能是在比较现代的数字通信方式与传统的打字或书写方式的差异。&lt;/think&gt;</b><br/><b>&lt;translate&gt; “如今, 人们可以在电脑屏幕上编写消息, 而无需接近任何针状物。”&lt;/translate&gt;</b></p>  |

Figure 6: Examples illustrating language-of-thought phenomenon, i.e., transition of the internal reasoning language in MT-R1-Zero models. The reasoning language transits from English at Step 0 to target language at Step 1600, indicated by **bold** text across various OOD test pairs (Finding 3).

# ⚡ Technique : raisonnement (Wang et al., 2025)

**[Source Sentence]** Passing by the tower with her husband on the path to the gate she could feel the vibrant air humming round them from the louvred belfry in the circle of sound, and it matched the highly-charged mental atmosphere in which she was living.

**[Thought During Translation]** I start by identifying and translating the key terms from the English sentence: 'tower' as '塔', 'husband' as '丈夫', 'gate' as '大门', 'vibrant' as '充满活力的', 'humming' as '嗡嗡作响', 'louvred' as '有百叶窗的', 'belfry' as '钟楼', 'circle' as '圆圈', 'sound' as '声音', 'highly-charged' as '高度紧张的', 'mental' as '精神的', and 'atmosphere' as '氛围'.

With these translations in mind, I attempt my first translation: 经过塔时, 她和丈夫走在通往大门的路上, 能感受到有百叶窗的钟楼在声音的圆圈中发出的充满活力的空气在他们周围嗡嗡作响, 这和她所处的高度紧张的精神氛围相匹配。①

I reflect on this attempt and realize that while it is generally accurate, the phrasing could be improved for better fluency and clarity. The segment '有百叶窗的钟楼在声音的圆圈中发出的充满活力的空气' feels cumbersome and could be simplified. Additionally, '高度紧张的精神氛围' might not fully capture the nuance of 'highly-charged mental atmosphere.'

I revise my translation: 经过塔时, 她和丈夫走在通往大门的路上, 感受到从有百叶窗的钟楼传来的充满活力的空气在他们周围轻轻回响, 这和她所处的高度活跃的精神氛围相契合。②

This revision feels more fluent and natural, effectively conveying the original meaning. The phrase '轻轻回响' improves the flow and captures the sense of the air humming around them. The term '高度活跃的精神氛围' seems a better fit for 'highly-charged mental atmosphere,' making the translation more accurate and easier to understand for a Chinese reader.

I make another revision: 经过塔时, 她和丈夫走在通往大门的路上, 感受到从有百叶窗的钟楼传来的充满活力的气息在他们周围轻轻回荡, 这和她内心的高度活跃氛围相契合。③

This version is very smooth and natural, effectively conveying the original meaning. The use of '轻轻回荡' is particularly apt, capturing the sense of the air humming around them. The phrase '高度活跃的精神氛围' accurately reflects the 'highly-charged mental atmosphere,' making the translation both precise and fluent. There are no significant issues that would hinder understanding for a Chinese reader.

Finally, I decide on the final translation: 经过塔时, 她和丈夫走在通往大门的路上, 感受到从有百叶窗的钟楼传来的充满活力的气息在他们周围轻轻回荡, 这和她内心的高度活跃氛围相映成趣。④

# Utilisation des LLM pour l'évaluation (Kocmi et Federmann, 2023)

Score the following translation from {source\_lang} to {target\_lang} **with respect to the human reference** on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

```
{source_lang} source: "{source_seg}"  
{target_lang} human reference: {reference_seg}  
{target_lang} translation: "{target_seg}"  
Score:
```

## ⊕ Ressources complémentaires sur les LLM en traduction

<https://github.com/hsing-wang/Awesome-LLM-MT>

# Bibliographie

- Agarwal, R., Vieillard, N., Stanczyk, P., Ramos, S., Geist, M., et Bachem, O. (juin 2023). *GKD: Generalized Knowledge Distillation for Auto-regressive Sequence Models*. [10.48550/arXiv.2306.13649](#)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (juillet 2020). *Language Models Are Few-Shot Learners* (Numéro arXiv:2005.14165). arXiv. [10.48550/arXiv.2005.14165](#)
- Chen, P., Guo, Z., Haddow, B., et Heafield, K. (mai 2024). *Iterative Translation Refinement with Large Language Models* (Numéro arXiv:2306.03856). arXiv. [10.48550/arXiv.2306.03856](#)
- Chen, R., Li, F., et Wang, Z. (2023). Prefix-LSDPM: A Few-shot Oriented Online Learning Session Dropout Prediction Model. *Journal of East China University of Science and Technology*, 49(5), 754-763. [10.14135/j.cnki.1006-3080.20230206003](#)
- Feng, Z., Zhang, Y., Li, H., Wu, B., Liao, J., Liu, W., Lang, J., Feng, Y., Wu, J., et Liu, Z. (avril 2025). TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement. In L. Chiruzzo, A. Ritter, et L. Wang (éds.), *Findings of the Association for Computational Linguistics: NAACL 2025: Findings of the Association for Computational Linguistics: NAACL 2025*. [10.18653/v1/2025.findings-naacl.218](#)
- Kocmi, T., et Federmann, C. (juin 2023). Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In M. Nurminen, J. Brenner, M. Koponen, S. Lattomaa, M. Mikhailov, F. Schierl, T. Ransinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, et H. Moniz (éds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation: Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
- Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press. [10.1017/CB09780511815829](#)
- Li, X. L., et Liang, P. (janvier 2021). *Prefix-Tuning: Optimizing Continuous Prompts for Generation* (Numéro arXiv:2101.00190). arXiv. [10.48550/arXiv.2101.00190](#)

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., Aji, A. F., Wong, D. F., Liu, S., et Wang, L. (avril 2024). *A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models* (Numéro arXiv:2305.01181). arXiv. [10.48550/arXiv.2305.01181](#)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., et Liu, P. J. (juillet 2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*.

Wang, J., Meng, F., Liang, Y., et Zhou, J. (février 2025). *DRT: Deep Reasoning Translation via Long Chain-of-Thought* (Numéro arXiv:2412.17498). arXiv. [10.48550/arXiv.2412.17498](#)

Zebaze, A., Sagot, B., et Bawden, R. (mars 2025). *Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation* (Numéro arXiv:2503.04554). arXiv. [10.48550/arXiv.2503.04554](#)