



TRANSLATING WITH LLMs

Guillaume Wisniewski

guillaume.wisniewski@u-paris.fr

UPCité Master in Computational Linguistics

November, 2025



What is a LLM?

2/49

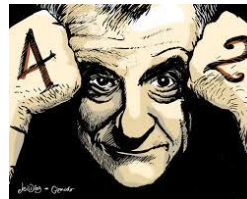


A BIASED DEFINITION



1. a transformer architecture (those thinking RNN are asked to leave the room quickly 🙄)
2. used to predict masked word
↳ self-supervised training
3. with a large number of parameters (and large training set)

3/49



Why is size so important?

4/49

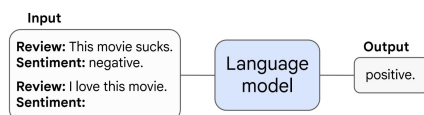


FEW-SHOT LEARNING

- GPT-3 has “capacities” that were not expected:
few-shot learning
- can be trained to do new tasks
↳ only by providing 2-3 examples
↳ directly in the prompt



EXAMPLE

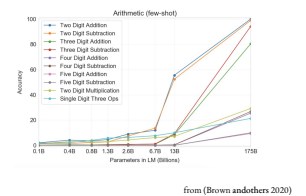


5/49



EMERGENT ABILITIES: E.G. ARITHMETIC

- GPT-3 knows how to add two numbers
↳ smaller models do not
- increasing model size $\Rightarrow$ new capacities
↳ emergence





IN A NUTSHELL

- phrase your task as an instruction...
- ...an LLM will solve it...
- ...without you having to program it explicitly



7/49



A PROMPTING APPROACH TO MT

A SIMPLE PROMPT FOR MT:

Source: $[X_1]$
Target: $[Y_1]$
Source: $[X_2]$
Target: $[Y_2]$
:
Source: $[X_n]$
Target: $[Y_n]$
Source: $[X]$
Target:

- ↪ Source and Target: language name
- ↪ several selection strategies (e.g. k -nn on the source representations)

- experiments described in (Vilar and others 2023)
- ↪ many similar papers (using different LLMs): (Agrawal and others 2022), (Zhang, Haddow and Birch 2023), (Hendy and others 2023) (*inter alia*)

8/49



RESULTS

LP	System	BLEURT ↑	BLEU ↑
en → fr	Google Trans.	76.5	45.7
	WMT-full random	75.9	42.3
	WMT-full $\hat{K}$ NN	75.3	41.8
	WMT-dev random	75.4	41.9
	high-end random	75.2	38.6
fr → en	Google Trans.	77.7	43.2
	WMT-full random	77.7	42.7
	WMT-full $\hat{K}$ NN	77.3	41.2
	WMT-dev random	77.2	42.1
	high-end random	77.6	40.4

(c) French→English (mt2014).

- results are comparable to that of an NMT system (on many languages)
- but much more expensive to generate (by any measure)

9/49



BE CAREFUL

Source Text: ff
LLM Output: This is an emoji and doesn't need translation. It's universally understood. ff
Source Text: Oooh, thank you "so" much!! I really appreciate this!! A
LLM Output: This can be translated to: 哦，太感谢了！！我真的非常感谢！！ A
Or a more natural and less literal translation: 哇，太感谢啦！！
我是超了级喜欢你！！ A
Source Text: I mean I've drunk all the beer but still don't have a job. WTF. How does this universe even work.
LLM Output: 我意思是，我把啤酒都喝光了，但还没有工作。搞什么鬼。这宇宙到底是怎么运作的？
Please note: that "drunk" is a typo and should be "drunk".

- LLMs answers contain more than translations
 - ↪ explanations, text to maximise users' engagement, ...
- maybe difficult to extract translations
 - ↪ e.g. Briakou and others 2024: extracting translation correctly change the ranking of systems by automatic metrics
- how to solve that?

10/49



BE CAREFUL

Source Text: ff
LLM Output: This is an emoji and doesn't need translation. It's universally understood. ff
Source Text: Oooh, thank you "so" much!! I really appreciate this!! A
LLM Output: This can be translated to: 哦，太感谢了！！我真的非常感谢！！ A
Or a more natural and less literal translation: 哇，太感谢啦！！
我是超了级喜欢你！！ A
Source Text: I mean I've drunk all the beer but still don't have a job. WTF. How does this universe even work.
LLM Output: 我意思是，我把啤酒都喝光了，但还没有工作。搞什么鬼。这宇宙到底是怎么运作的？
Please note: that "drunk" is a typo and should be "drunk".

- LLMs answers contain more than translations
 - ↪ explanations, text to maximise users' engagement, ...
- maybe difficult to extract translations
 - ↪ e.g. Briakou and others 2024: extracting translation correctly change the ranking of systems by automatic metrics
- how to solve that?
 - ↪ ask (politely) in the prompt
 - ↪ constrain output
 - ↪ do not use prompting

10/49



A PARADIGME SHIFT FOR MT

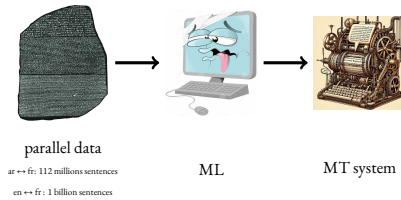


LLMs translate in fundamentally different ways from statistical and NMT systems 🤖🧠🤖

11/49



STATISTICAL MACHINE TRANSLATION



- statistical learning (first: phrase-based, then NMT)
- ↳ learns to **reproduce** translations from a training set
- very good results when a large, in-domain, parallel corpus is available

12/49



BASICALLY...

MT systems:

- learn to **imit/reproduce** the translations seen during training
- ↳ identify patterns / co-occurrences
- can only translate words it has observed
- ↳ to translate “traduction automatique”, the term must appear in a (many) parallel sentences with the correct translation
- ↳ only a distorted image of the reality



13/49



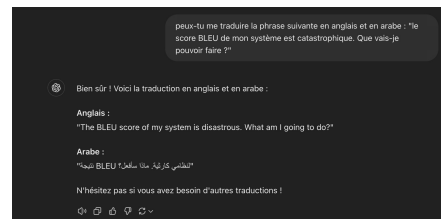
WHY IS TRANSLATION WITH A LLM DIFFERENT?

- LLMs are trained from monolingual data only
- ↳ even if parallel corpora are included in monolingual data
⇒ no explicit cross-lingual supervision
- translation capacities = emerges

14/49



TRANSLATING WITH PROMPTS



- as DEEPL
- but without parallel data !

15/49



TRANSLATION WITH LLMs IS MAGIC ©



- translation without parallel data
- ↳ no more domain / language problems
- controls style, terminology, length, ... using the prompt
- ↳ “translates the following text into formal French...”
- ↳ “translates the following text knowing that A translates into B, C into D, ...”
- ↳ “Here are some examples of translations from the [...] domain translate me the following text”

16/49



TERMINOLOGY EXTRACTION



Please extract the technical terms from the following text and list them in a table in Arabic, English and French. [Wikipedia page on Translation Studies]

English	Arabic	French
Sense-for-sense translation	ترجمة بمعنى مقابل	traduction sens pour sens
Prescriptive	اقتضائي	prescriptif
Descriptive	وصفي	descriptif
Linguistic transformation	التحويل اللغوي	transformation linguistique
Typology	التصنيف	typologie
Cultural turn	التحول الثقافي	tournant culturel
Polysystem theory	نظرية النظم المتعددة	théorie des polysystèmes
Manipulation	تلاعب	manipulation
Patronage	رعاية	mécénat
Skopos theory	نظرية سكوبوس	théorie du Skopos
Translation history	تاريخ الترجمة	histoire de la traduction
Sociology of translation	علم اجتماع الترجمة	sociologie de la traduction
Post-colonial translation	الترجمة بعد الاستعمار	traduction postcoloniale
Gender studies	دراسات الجنس	études de genre
Audiovisual translation	الترجمة في الوسائط المتعددة	traduction audiovisuelle

17/49



EXPLIQUER LES TRADUCTIONS

dans la traduction anglaise, pourquoi n'as-tu pas traduit « bleu » en « blue »

Excellente question ! En effet, dans ce contexte, BLEU n'est pas le mot français pour "bleu" (la couleur) mais une abréviation pour "Bilingual Evaluation Understudy". C'est une métrique utilisée pour évaluer la qualité des traductions automatiques en comparant les résultats de la machine avec des traductions de référence humaines.

BLEU est donc un terme technique qui ne se traduit pas en "bleu" en anglais car il s'agit d'un acronyme internationalement reconnu dans le domaine de la traduction automatique.

18/49



IDENTIFYING TRANSLATION DIFFICULTIES

- Je suis un traducteur professionnel qui doit traduire la poésie suivante : « Mon enfant, ma sœur, Songe à la douceur, D'aller là-bas vivre ensemble ! [...] ». Quelles seront les difficultés et les points de vigilance
- Traduire cette poésie, tirée de « L'Invitation au voyage » de Charles Baudelaire, pose plusieurs défis en raison de la densité émotionnelle, des sonorités, et des subtilités de sens qui caractérisent le poème. Voici quelques points clés et difficultés potentielles auxquelles prêter attention : [...] « Mon enfant, ma sœur » crée immédiatement une relation intime et tendre, où le locuteur invite une personne proche dans une sorte de rêve commun. Cette familiarité douce peut être difficile à rendre en anglais, où les expressions intimes n'ont pas toujours le même naturel ou la même portée. [...] Par exemple, l'idée d'« aimer à loisir, aimer et mourir » pourrait être interprétée comme une célébration de l'amour dans sa liberté totale, un amour si complet qu'il contient même la mort. Cette idée doit se transmettre avec la même intensité et ambiguïté en anglais, ce qui peut nécessiter des réajustements créatifs.

19/49



EVALUATE TRANSLATIONS (1)

- use an LLM to identify errors in a translation
- providing it with a typology of errors

Tâche : annoter une traduction
Objectif : repérer des erreurs sur la base d'une typologie d'erreurs que je te fournis.
Type de texte : résumé d'article scientifique dans le domaine du TAL
Fichier joint : MANUEL D'ANNOTATION, qui contient des explications plus détaillées et des exemples des types d'erreurs que je vais te fournir ci-dessous.
Présentation de la sortie :
- 1re phrase source
- 1re phrase cible dans la traduction
- liste les erreurs (mets les erreurs en gras, pour les mettre en évidence)
- 2e phrase source
- 2e phrase cible dans la traduction
Note les erreurs (mets les erreurs en gras, pour les mettre en

20/49



EVALUATE TRANSLATIONS (2)

Typologie d'erreurs à suivre méticuleusement : veille à utiliser les types d'erreurs présents et n'en invente aucun. De même, respecte les codes liés à chaque type d'erreur à la lettre ; ne prends donc aucune liberté.
Explication de la typologie : elle est divisée en 3 grandes catégories d'erreurs : les erreurs de transfert de contenu (erreurs altérant le sens du message ou entravant sa compréhension), les erreurs de langue, et les erreurs liées aux outils ou à leur maîtrise.
Voici la typologie :
1. Transfert-contenu (GRANDE CATÉGORIE, NE PAS UTILISER)
1.1. Omission_TR-OM
* Une omission se produit lorsqu'il manque, dans la traduction, une idée qui est présente dans le texte source. Il ne faut pas confondre

21/49



EVALUATE TRANSLATIONS (3)

Phrase source 1 :
INCLURE: a Dataset and Toolkit for Inclusive French Translation

Phrase cible 1 :
Inclure un ensemble de données et d'outils pour la traduction du français inclusif

Erreurs :

- Mauvaise-casse_LA-HY-CA : Inclure devrait être en majuscule, car il s'agit du nom d'un projet ou ensemble de données (INCLURE).
- Choix-incorrec-Termino_LA-TL-INS : un ensemble de données est correct, mais Dataset pourrait se traduire par corpus pour correspondre au contexte scientifique de la phrase.

22/49



EVALUATE TRANSLATIONS (4)

Phrase source 2 :
Inclusive French (gender-neutral language) is a variety of French that is used to highlight awareness of gender and identity against Standard French, which enforces the use of masculine for generic usage or plural.

Phrase cible 2 :
Le français inclusif (langue sexuellement neutre) est une variété de français utilisée pour mettre en évidence la sensibilisation au genre et à l'identité par rapport au français standard, qui impose l'utilisation du masculin pour l'usage générique ou du pluriel.

Erreurs :

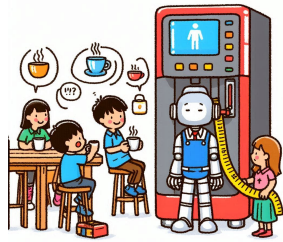
- Choix-incorrec-Termino_LA-TL-ING : langue sexuellement neutre est une traduction incorrecte de gender-neutral language. La formulation correcte serait langue neutre en termes de genre.
- Trop-littérale_TR-SI-TL : mettre en évidence la sensibilisation au genre et à l'identité est

23/49



WARNING

- these are just examples of what can be done (or, more precisely, what has been done)
- ↳ it's up to you to imagine new uses
- today main challenges
- ↳ true evaluation of llms capabilities
- ↳ understand if and when it fails



24/49



TO BE CLEARER



- no explanation for LLMs' cross-lingual capabilities (as for many other capacities)
- ↳ to be more precise: contradictory and partial explanations
- for the moment just a very (too) large number of observations

25/49

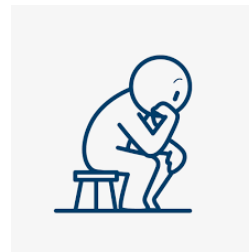


TWO TECHNIQUES TO IMPROVE LLM CAPACITY TO REASON



- instruction fine-tuning
- chain-of-thought prompting

26/49



Instruction Fine-Tuning

27/49



INSTRUCTION FINE-TUNING

- “instruction tuning”: fine-tuning a LM on a collection of tasks described by instructions
- improves the zero-shot performance of LM
- ↳ no need to explain the task at hand, the model already knows it!



28/49



RECALL: FINE-TUNING A DECODER

29/49



KISS

Keep It Simple Stupid.

THE no robots CORPUS

prompt Why is Aberdeen called the Granite City?

answer Aberdeen's nickname, the Granite City, comes from the city's history and contribution of granite, in addition to the architectural style of more than fifty percent of the city's buildings. Many of the city's most popular buildings and residential properties were built using granite from large quarries around the North-East, most notably the Rubislaw Quarry in Aberdeen's west end. As technology and practices for handling a large amount of rock improved, Aberdeen quickly became the granite capital of the world, exporting granite by sea to several areas globally.

THE no robots CORPUS

prompt I have 2 lists below. Can you show me how to make a bar chart from them using Python?

```
1 list1 = ["USA", "China", "Russia", "Brazil", "Japan"]
2 list2 = [22, 53, 32, 45, 56]
```

answer There is a useful module called Matplotlib that can help you achieve this. If you have not already, you will need to install the module first. To install it with PIP, open your command line and enter:

```
pip install matplotlib
```

To install it with Conda, enter:

```
conda install -c conda-forge matplotlib
```

Then you can import the module into your Python code. Here is the full code:

```
1 from matplotlib import pyplot as plt
2
3 list1 = ["USA", "China", "Russia", "Brazil", "Japan"]
4 list2 = [22, 53, 32, 45, 56]
5
6 plt.bar(list1, list2)\nplt.show()
```

Different categories of prompts:

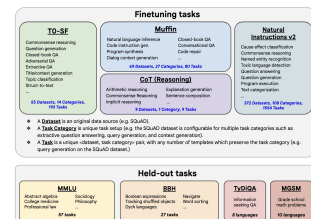
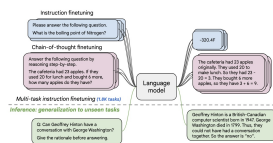
- Generation: 45.7%
- OpenQA: 12.4%
- Brainstorm: 11.2%
- Chat: 8.4%
- Rewrite: 6.6%
- Summarize: 4.2%
- Classify: 3.5%
- Coding
- ...

↪ 10k examples

↪ several other datasets: Dolly-15k, ...



- instruction fine-tuning: improve the model performance on new tasks 🤖
- ↳ i.e. tasks that have not been seen during training!
- See: (Wei, Bosma and others 2022) or (Chung and others 2024) more generally: look for the work of Jason Wei





THE RESULTS

	MMLU	BBH-nlp	BBH-ulg	TyDiQA	MGSM
Prior best	69.3 ^a	73.5 ^a	73.9 ^a	81.9 ^a	55.0 ^a
PaLM 540B					
- direct prompting	69.3	62.7	38.3	52.9	18.3
- CoT prompting	64.5	71.2	57.6	-	45.9
- CoT + self-consistency	69.5	78.2	62.2	-	57.9
Flan-PaLM 540B					
- direct prompting	72.2	70.0	48.2	67.8	21.2
- CoT prompting	70.2	72.4	61.3	-	57.0
- CoT + self-consistency	75.2	78.4	66.5	-	72.0

Table 4: Flan-PaLM outperforms PaLM on all evaluation benchmarks. Prior best are the following: a: PaLM without CoT prompting (Chowdhery et al., 2022). b: Codex with CoT prompting but no self-consistency (code-davinci-002; Chen et al., 2021). c: Finetuned ByT5 (Xue et al., 2022). d: PaLM + Google translate API with CoT prompting but no self-consistency (Shi et al., 2022). The MMLU results are on the test set.

from (Chung and others 2024)

36/49



ANOTHER WONDERFUL TECHNIQUE



What is Chain-of-Thought prompting?

37/49



CoT PROMPTING

- guiding LLM to generate a **step-by-step sequence of reasoning processes** rather than jumping directly to an answer
- identify (and solve!) subtasks to solve a problem
- “unlock” reasoning capabilities in LLMs through structured prompting

38/49



A PICTURE IS WORTH A THOUSAND WORDS

Standard Prompting

Model Input

Q. Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A. The answer is 11.

Q. The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A. The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q. Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A. Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q. The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

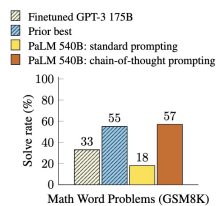
A. The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

from (Wei, Wang and others 2024)

39/49



IMPACT



- impressive results
- again: (almost) no explanation

40/49



PROPERTIES OF CoT

- Decomposition of Multi-Step Problems
- Interpretability of Model Behavior
- Versatility Across Tasks $\Rightarrow$ can be used in any task that humans solve using language-based reasoning.
- Ease of Elicitation in LLMs $\Rightarrow$ can be “triggered” using few-shot learning

41/49



EN METTANT TOUT BOUT À BOUT

Pre-writing Research	Outlining	Revising	Proofreading
The first step in the translation process is to understand the source text. This involves identifying the main purpose of the text, the target audience, and the context in which it will be used. It is important to read the text carefully and take notes on any unfamiliar words or phrases. Once you have a good understanding of the source text, you can begin to plan your translation. This involves deciding on a strategy for dealing with any cultural differences or idiomatic expressions. It is also important to think about the layout and formatting of the translated text. After you have planned your translation, you can begin to write the first draft. This is a good time to use the notes you took during your research and to refer back to the source text as you write. It is important to write clearly and to use simple words where possible. Once you have written the first draft, you should leave it for a while before revising it. This will allow you to see the text with fresh eyes and to identify any areas that need improvement. You should also ask someone else to read your translation and give you feedback. Finally, you should proofread your translation carefully before you submit it. This involves checking for any spelling or grammar errors and making sure that the text is easy to read and understand.	Once you have a good understanding of the source text, you can begin to plan your translation. This involves deciding on a strategy for dealing with any cultural differences or idiomatic expressions. It is also important to think about the layout and formatting of the translated text. After you have planned your translation, you can begin to write the first draft. This is a good time to use the notes you took during your research and to refer back to the source text as you write. It is important to write clearly and to use simple words where possible. Once you have written the first draft, you should leave it for a while before revising it. This will allow you to see the text with fresh eyes and to identify any areas that need improvement. You should also ask someone else to read your translation and give you feedback. Finally, you should proofread your translation carefully before you submit it. This involves checking for any spelling or grammar errors and making sure that the text is easy to read and understand.	Once you have a good understanding of the source text, you can begin to plan your translation. This involves deciding on a strategy for dealing with any cultural differences or idiomatic expressions. It is also important to think about the layout and formatting of the translated text. After you have planned your translation, you can begin to write the first draft. This is a good time to use the notes you took during your research and to refer back to the source text as you write. It is important to write clearly and to use simple words where possible. Once you have written the first draft, you should leave it for a while before revising it. This will allow you to see the text with fresh eyes and to identify any areas that need improvement. You should also ask someone else to read your translation and give you feedback. Finally, you should proofread your translation carefully before you submit it. This involves checking for any spelling or grammar errors and making sure that the text is easy to read and understand.	Once you have a good understanding of the source text, you can begin to plan your translation. This involves deciding on a strategy for dealing with any cultural differences or idiomatic expressions. It is also important to think about the layout and formatting of the translated text. After you have planned your translation, you can begin to write the first draft. This is a good time to use the notes you took during your research and to refer back to the source text as you write. It is important to write clearly and to use simple words where possible. Once you have written the first draft, you should leave it for a while before revising it. This will allow you to see the text with fresh eyes and to identify any areas that need improvement. You should also ask someone else to read your translation and give you feedback. Finally, you should proofread your translation carefully before you submit it. This involves checking for any spelling or grammar errors and making sure that the text is easy to read and understand.

Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts, E. Briakou, J. Luo, C. Cherry, M. Freitag, arXiv:2409.06790, 2024

⇒ améliore les performances (meilleurs scores WMT'24)

42/49



43/49



PROMPTING

- prompting used to be a big thinks
 - ↳ “prompt engineer” the new dream job
 - ↳ so many books, tutorial, training programs, ...
- no longer true !
 - ↳ LLMs are so capable they grasp what you want better than you do.

44/49



PROMPT REWRITING

- most LLMs rewrite your prompts
 - ↳ ensure output consistency (e.g. ollama json schema)
 - ↳ ensure better interaction/answer

45/49



THE LAST TRICK(S) TO KILL PROMPTING



- “Let’s think this step by step”
 - ↳ (kojima22large2025) “better” than all CoT strategies
- “Generate 5 responses to this prompt, each with their probability.”
 - ↳ (zhang2025verbalized) (published 1st Oct. 2025)
 - ↳ “verbalized sampling”
 - ↳ increase diversity/creativity (reduce typicality)

46/49



Part I

REFERENCES

47/49



-  Agrawal, Sweta and others (2022). *In-context Examples Selection for Machine Translation*.
-  Briakou, Eleftheria and others (2024). *On the Implications of Verbose LLM Outputs: A Case Study in Translation Evaluation*.
-  Brown, Tom and others (2020). "Language Models are Few-Shot Learners". in *Advances in Neural Information Processing Systems*: by editor H. Larochelle and others. volume 33. Curran Associates, Inc., pages 1877–1901.
-  Chung, Hyung Won and others (2024). "Scaling Instruction-Finetuned Language Models". in *Journal of Machine Learning Research*: 25.70, pages 1–53.
-  Hendy, Amr and others (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*.



-  Vilar, David and others (July 2023). "Prompting PaLM for Translation: Assessing Strategies and Performance". in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: by editor Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pages 15406–15427.
-  Wei, Jason, Maarten Paul Bosma and others (2022). "Finetuned Language Models are Zero-Shot Learners". in.
-  Wei, Jason, Xuezhi Wang and others (2024). "Chain-of-thought prompting elicits reasoning in large language models". in *Proceedings of the 36th International Conference on Neural Information Processing Systems: NIPS '22*. New Orleans, LA, USA: Curran Associates Inc.
-  Zhang, Biao, Barry Haddow and Alexandra Birch (2023). *Prompting Large Language Model for Machine Translation: A Case Study*.