Chain-of-Thought Reasoning Improves Context-Aware Translation with Large Language Models

Shabnam Ataee[1], **Hugo Huart**[1], **Andrei Popescu-Belis**[1,2]
[1]HEIG-VD/HES-SO, Route de Cheseaux 1, 1401 Yverdon-les-Bains, Switzerland
[2]EPFL, 1015 Lausanne, Switzerland
{shabnam.ataee, hugo.huart, andrei.popescu-belis}@heig-vd.ch

Abstract

This paper assesses the ability of large language models (LLMs) to translate texts that include inter-sentential dependencies. We use the English-French DiscEvalMT benchmark (Bawden et al., 2018) with pairs of sentences containing translation challenges for pronominal anaphora and lexical cohesion. We evaluate 12 LLMs from the DeepSeek-R1, GPT, Llama, Mistral and Phi families on two tasks: (1) distinguish a correct translation from a wrong but plausible one; and (2) generate a correct translation. We compare prompts that encourage chain-of-thought reasoning with those that do not. The best models take advantage of reasoning and reach about 90% accuracy on the first task and COMET scores of about 92% on the second task, with GPT-4, GPT-4o and Phi standing out. Moreover, we observe a "wise get wiser" effect: the improvements through reasoning are larger for models that already perform well without reasoning.

Keywords: Context-aware MT, Translation with LLMs, Chain-of-thought reasoning for translation.

1. Introduction

Large language models (LLMs) have shown impressive capacities for translation, among many other tasks, but their translations are still not perfect (Kocmi et al., 2024; Cui et al., 2025). To increase quality, we examine whether chain-of-thought (CoT) reasoning helps to improve referential and lexical cohesion across sentences. Indeed, when prompted to generate a step-by-step explanation of their answers, LLMs improve substantially on many tasks (Wei et al., 2022).

We hypothesize that CoT can improve the translation of elements that maintain coherence across sentences by making explicit translation decisions. To test our hypothesis, we use the DiscEvalMT English-French test suites for pronominal anaphora and lexical consistency designed by Bawden et al. (2018), in two settings. First, we ask an LLM to select the correct translation among the two contrastive alternatives present in the data. Second, we ask the LLM to translate each test sentence and we score the result against a reference translation. In both settings, we demonstrate the advantages of CoT prompting, though only for the largest LLMs, which improve their performance when encouraged to reason before translating.

The paper is organized as follows. We review related work in Section 2. We present the benchmark data and the evaluation metrics in Section 3. The tested LLMs are listed in Section 4: they include 4 GPT proprietary models from OpenAI and 8 open-weight models. The prompts and the results for the contrastive evaluation task (i.e., selecting the correct translation) are presented in Section 5, while those for the translation task are in Section 6. Our main contributions are the following ones:

- We assess the ability of 12 LLMs to translate coherently, showing that they reach a new state of the art on the DiscEvalMT benchmark for pronominal anaphora and lexical cohesion.
- We show that the translation task is reliably scored, as BLEU, chrF, BERTscore, and COMET scores are correlated.
- We demonstrate that CoT prompting improves coherence, though only for the best models. In other words, powerful models have a better capacity to leverage reasoning, which we call a "wise get wiser" effect.
- We share the outputs of the tested systems at https://zenodo.org/records/17483104.

2. Related Work

The translation capabilities of multilingual LLMs have attracted growing attention since 2023. Conversational LLMs are now often used for translation tasks, which represent around 4.5% of non-professional interactions with ChatGPT (Chatterji et al., 2025). Therefore, studies evaluating the translation capability of LLMs are of high importance, all the more that translation is included in common benchmarks for LLMs. LLMs have been tested since 2023 at the Workshops on Machine Translation (Kocmi et al., 2023, 2024), and in 2025, the majority of submissions to the general task were LLM-based (Kocmi et al., 2025). At the same time,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3725–3741
11–16 May 2026.

the first studies of LLMs for MT included quantitative evaluations (Vilar et al., 2023; Zhang et al., 2023; Hendy et al., 2023). Some studies focused on evaluation only: for instance, Bawden and Yvon (2023) demonstrated the strong MT performance of BLOOM, particularly in few-shot settings and for high-resource language pairs. Fine-tuning of LLMs for document-level MT was found to be moderately beneficial (Wu et al., 2024).

Word-sense disambiguation (WSD) for translation is one of the challenges on which LLMs have been evaluated. Test suites related to WSD include DiscEvalMT (Bawden et al., 2018) which we use in this paper and describe below, ContraWSD (Rios Gonzales et al., 2017) with about 7,000 sentences with reference translations, and a smaller test suite to score word translations (Rios et al., 2018). The DiBiMT benchmark with about 600 examples, for translation from English to five other languages, was proposed by Campolungo et al. (2022) but remains private and systems must be submitted to its owners. DiBiMT was used by Iyer et al. (2023) to score the capacity of LLMs for WSD.

The capacity to translate potentially ambiguous pronouns can be tested using contrastive test suites such as DiscEvalMT (Bawden et al., 2018), ContraPRO (Müller et al., 2018; Lopes et al., 2020), or PROTEST (Guillou and Hardmeier, 2016), which was used at WMT 2018 (Guillou et al., 2018). A test suite for ellipses and lexical cohesion was designed by Voita et al. (2019). A review of scores reached on the DiscEvalMT data by various context-aware MT systems appears below in Section 5.2. An alternative approach is to align and compare pronouns in the MT output with those in a reference translation (Miculicich Werlen and Popescu-Belis, 2017), but scores may be biased by the choice of the antecedent.

The correct translation of pronouns or ambiguous words is often related to the more general ability to leverage inter-sentential dependencies and generate a contextually-correct translation, reviewed in several studies (Popescu-Belis, 2019; Maruf et al., 2021; Jin et al., 2023; Castilho and Knowles, 2025). Document-level translation and its evaluation are important because MT outputs may appear competitive with human translations at the sentence level, but not at the document level (Läubli et al., 2018). An evaluation of GPT-3.5 and GPT-4 on document-level translation (Wang et al., 2023) found that they surpassed non-LLM systems in human ratings. Karpinska and Iyyer (2023) found that LLMs translate better entire paragraphs than individual sentences, in the case of literary translation. In an analysis of the translation of six types of contextual dependencies (with one sentence per type), Castilho et al. (2023) found that GPT 3.5 outperformed NMT systems on three high-resource languages, but not on a low-resource one (Gaelic). Beyond direct translation, interaction was a key element in the WSD process proposed by Pilault et al. (2023), using questions and answers as a chain-of-thought input to an LLM.

Recent work has explored how reasoning can improve LLM-based translation, inspired by the seminal work of Wei et al. (2022), who introduced chain-of-thought (CoT) prompting and showed that guiding models through intermediate steps improves performance on complex tasks. Liu et al. (2025) argue that reasoning enhances translation by improving coherence, cultural alignment, and self-reflection. He et al. (2025) present R1-T1, which uses expert CoT templates and reinforcement learning to encourage inference-time reasoning for translation. Ye et al. (2025) found that reasoning models outperform standard LLMs on semantically complex and domain-specific MT, especially for longer texts.

From the variety of MT evaluation metrics, many have also been considered for extension to document-level evaluation (Dahan et al., 2024). Given that we only consider here sentence pairs, we will consider four widely used metrics. We will use BLEU and chrF (Papineni et al., 2002; Popović, 2015), which are based on surface-level overlap, although they offer limited insights into discourse-level quality. We will also use embedding-based metrics such as BERTScore (Zhang et al., 2020), and trained metrics such as COMET (Rei et al., 2020), as they model semantic similarity and are now frequently used.

LLMs have also been used for MT evaluation: Kocmi and Federmann (2023) showed that GPT-based metrics reach state-of-the-art correlations with human judgments on WMT benchmarks. The benefits of LLM reasoning for this task are still under investigation: Larionov et al. (2025) found that reasoning-enabled LLMs such as DeepSeek-R1 and OpenAI's o3-mini do not always outperform their non-reasoning counterparts in terms of alignment with human judgments of MT quality.

3. Evaluation Data and Metrics

3.1. Benchmark Data

This study uses the DiscEvalMT benchmark (Bawden, 2018; Bawden et al., 2018), designed to evaluate challenges in English-to-French translation that arise from two types of inter-sentential dependencies: pronominal anaphora and lexical cohesion.[1] The test cases are manually constructed. Each test item consists of two English sentences ('context' and 'current'), a French translation of the first one, and two alternative French translations of the

[1] https://github.com/rbawden

	Anaphora	Lexical choice
EN context	It's been a while since I last went to the *river*.	And then the *attack* took place.
EN current	It feels great to finally see *it*.	A truly terrible *attack*.
FR context (given)	Ça fait longtemps que je n'ai pas été à la *rivière*.	Et puis l'*attaque* a eu lieu.
FR current #1 (correct)	C'est chouette de *la* voir enfin.	Une *attaque* vraiment affreuse.
FR current #2 (incorrect)	C'est chouette de *le* voir enfin.	Un *assaut* vraiment affreux.

Table 1: Contrastive test items for anaphora and lexical choice from DiscEvalMT (Bawden et al., 2018).

second one: one is contextually appropriate, preserving coherence across the two sentences, while the other introduces a discourse-level error. Together, they form a contrastive pair among which a system should distinguish the correct translation.

An item for testing the translation of pronominal anaphora is shown in Table 1 (center). The word 'river' is translated as 'rivière' (feminine), so FR translation #1 has correct gender agreement between the object pronoun 'la' and the antecedent 'rivière', while the second one has wrong agreement. To control for chance, in another test item, 'river' is translated as 'fleuve' (masculine), making FR translation #2 correct. Two additional test items are made with approximate translations of the antecedent (e.g., 'piscine' (fem.) or 'cinéma' (masc.)). In these cases, the translation of the second sentence should still use a pronoun of the same gender as the antecedent, to maintain agreement.

For lexical choice, each item contains a word in the context sentence, which can be translated in several ways. The correct translation of the second sentence should use the same word as in the reference translation of the first sentence, to maintain consistency. In Table 1 (right), the word 'attack' is translated as 'attaque' in the context sentence, so the current sentence should reuse the same word, as in FR translation #1. In an additional test item, 'attack' is translated by 'assaut' in the context sentence, making FR translation #2 correct.

The anaphora test set includes 200 items (50 sets of four similar context sentences), each with two candidate translations of the second sentence. Similarly, the lexical choice test set includes 200 items with two candidates each. We selected the first half of the data for developing the prompts, and kept the second half for final testing.

3.2. Metrics

In the first part of our study (Section 5), the task of the LLM is to distinguish the correct translation of the second sentence in each pair from the wrong one (contrastive task). In the second part (Section 6) the task of the LLM is to translate the second sentence into French.

The contrastive task is scored by counting the number of times the correct translation was identified. The LLMs are prompted to output their response either as 'Choice: (1)' or as 'Choice: (2)'. This can be preceded by reasoning in case of CoT prompting. Our evaluation script is tolerant to minor formatting differences in the LLM's output: responses are accepted as correct as long as they clearly indicate the right option (1 or 2); if they cannot be parsed, they are counted as incorrect. The mean accuracy score is the average over all the answers.

To avoid systematic bias in favor of either candidate #1 or #2, we present each test item twice, once with the correct option being first, and once second. Accordingly, we measure consistency, i.e. the sensitivity of the model to the position of the correct translation among the two options. Ideally, the response should not depend on the position. We define inconsistency as the absolute difference between the accuracy score when the correct option is presented first ($\text{ACC}_{\text{correct=1}}$) and the accuracy score when it is presented second ($\text{ACC}_{\text{correct=2}}$), normalized by the sum of these accuracies. Inconsistency varies from 0, if accuracy is not influenced at all by the position of the correct answer, to 1, e.g. if an LLM always answers '1'.

$$\text{Inc.} = \frac{|\text{ACC}_{\text{correct=1}} - \text{ACC}_{\text{correct=2}}|}{\text{ACC}_{\text{correct=1}} + \text{ACC}_{\text{correct=2}}}$$

For the translation task, we provide the LLM with the two EN source sentences, as well as the FR translation of the context sentence, and ask it to translate the second EN sentence (possibly outputting reasoning). We parse the output to isolate the translation, and score it by comparing it to the correct translation using four metrics: BLEU, chrF, BERTscore and COMET. The first two capture surface-level overlap, while BERTScore and COMET assess semantic similarity using embeddings. All metrics are computed using their official Python implementations available via the `sacrebleu`[2], `bert-score`[3], and `unbabel-comet`[4] packages. BLEU and

[2] https://github.com/mjpost/sacrebleu
[3] https://github.com/Tiiiger/bert_score
[4] https://github.com/Unbabel/COMET

chrF are calculated with the default configurations of `sacrebleu`. BERTScore is computed using the default multilingual model (`bert-base-multilingual-cased`) and the language parameter set to French (`lang='fr'`). COMET uses the `Unbabel/wmt22-comet-da` model trained on human-annotated translations, which was shown to correlate strongly with human judgments. Given the high density of pronouns, we believe that scoring the entire sentence with frequently-used global metrics is a better solution than comparing pronouns with the APT metric (Miculicich Werlen and Popescu-Belis, 2017).

4. Evaluated LLMs and Baseline

We evaluated 12 LLMs across different model families and scales. We evaluated GPT models from OpenAI – GPT-3.5-turbo, GPT-4, GPT-4-turbo, and GPT-4o – accessed through API requests.[5] We evaluated the following open-weight models: Mistral (7.25B), Phi-4 (14.7B), three versions of Llama (3.1 with 8B, 3.2 with 3.2B, and 3.3 with 70B parameters), and three versions of DeepSeek-R1 (8B, 14B, and 32B); the latter are trained to reason by default, as discussed in Section 5 below. We do not report results with the Tower LLM (Alves et al., 2024), although it is a model that was fine-tuned specifically for translation. However, Tower was not instructed for CoT prompting, and we obtained low scores with it in preliminary experiments.

The LLMs were run locally using the `Ollama` framework,[6] which provides its own versions of the above models. For each model, we used the default quantized versions provided by Ollama (Q4_0 or Q4_K_M). Experiments were done on a Linux server with four NVIDIA RTX 2080 Ti GPUs with 11 GB VRAM each, using parallel querying to maximize throughput, with up to 16 simultaneous query threads. This limit was determined empirically to ensure stable and efficient throughput.

We include for comparison an encoder-decoder neural MT system, namely the distilled NLLB-200 multilingual model with 600M parameters (NLLB-Team et al., 2022). The model was queried using the Transformers library from Hugging Face.[7] For both tasks (contrastive and translation), we used prefix decoding: the French translation of the first sentence was added as a fixed prefix to the decoder. Then, for the contrastive task, we generate the translation of the second sentence, and compare it to each of the two options in the test pair using BERTScore. We assume that the most similar one is the model's answer. For the translation task, the MT model also uses prefix decoding to generate the translation of the second sentence, which is then evaluated.

5. Contrastive Task

5.1. Prompts: Reasoning or Not

Prompts to LLMs are typically made of a *system prompt*, which indicates the role, persona, or style expected for the answer, and a *user prompt*, which specifies the task, ending with data for the task. Here, we instruct the LLMs to solve the contrastive task, i.e. select the correct translation, with several variants for the system and user prompts.

The *system prompt* can be either *empty* (no system prompt), *simple*, or *detailed*. Each of these versions is shown in Appendix A.1. The detailed version is twice longer than the simple one and contains more constraints, but neither of them gives any instruction on how to solve the task. The system prompts are the same for the two benchmarks, anaphora or lexical cohesion, on the contrastive task.

The *user prompt* can be either a *simple* definition of the task, which includes the data, or instructions on how to reason *step-by-step* to solve it. These prompts are also shown in Appendix A.1. The step-by-step reasoning instructions for the anaphora and the lexical choice tasks are very similar: the main difference is at the third step. For anaphora, this is "Find the text in English line 1 to which the text found at Step 2 refers" while for lexical cohesion this is "Find the text in English line 1 which is identical to the text found at Step 2".[8]

We designed the prompts through a series of experiments on half of the benchmark data (the development set) with 12 LLMs. We settled on four combinations of system and user prompts: (1) no system prompt and simple user prompt; (2) simple system and user prompts; (3) detailed system prompt and simple user prompt; (4) simple system prompt and step-by-step (reasoning) user prompt.

The DeepSeek-R1 models are "reasoning" models, as they were trained to generate a detailed solution with multiple verifications and to conclude with the final answer. The usage instructions for DeepSeek-R1, which we follow, recommend to avoid a system prompt, but to include the words "please reason step by step" and indicate the explicit markup of the answer in the user prompt.

[5] https://platform.openai.com

[6] https://ollama.com

[7] https://huggingface.co/facebook/nllb-200-distilled-600M.

[8] These instructions are oriented towards solving the anaphora or lexical cohesion problems, and are therefore more focused than the document-level step-by-step translation considered by Briakou et al. (2024).

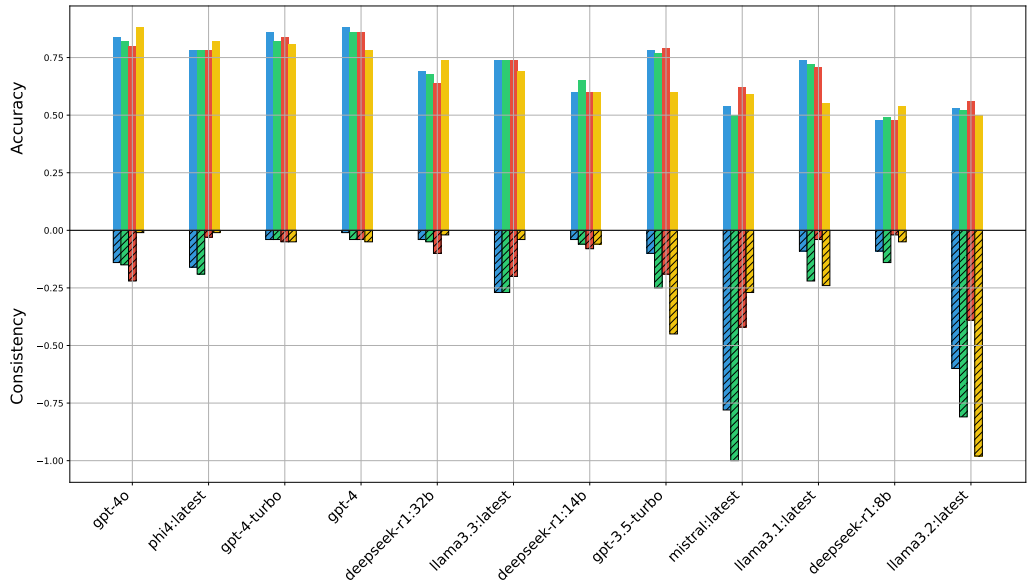

Figure 1: Mean accuracy (top bars) and inconsistency (bottom bars, smaller are better) on the contrastive anaphora task, validation set. For each LLM, the four bars are those of the prompts: (1) no system prompt and simple user prompt; (2) simple system and user prompts; (3) detailed system prompt and simple user prompt; (4) simple system prompt and step-by-step (reasoning) user prompt. LLMs are ranked by decreasing scores obtained with the last prompt (yellow), which reaches highest overall performance.

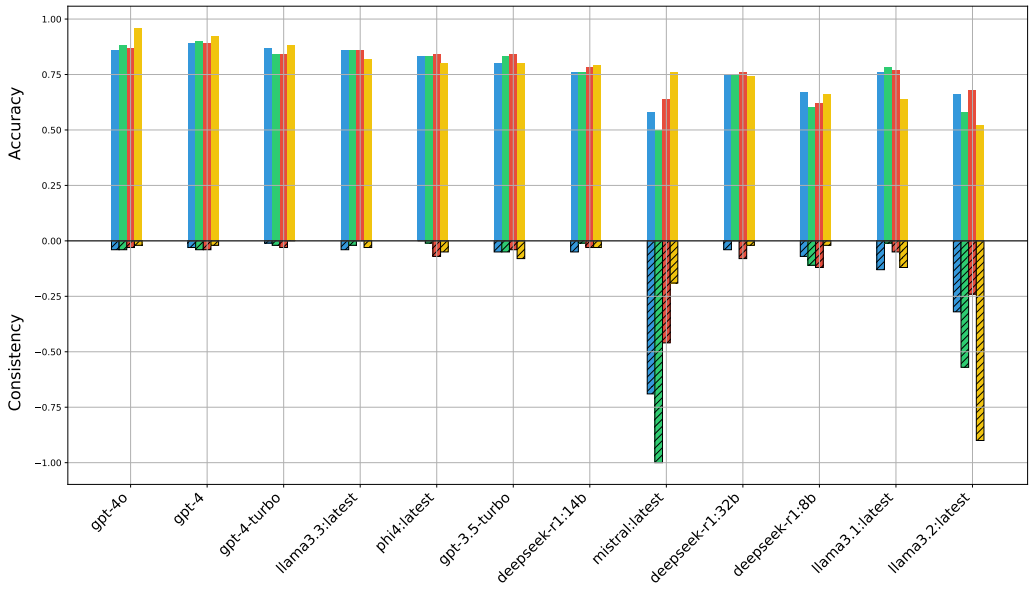

Figure 2: Mean accuracy (top bars) and inconsistency (bottom bars, smaller are better) across prompt configurations for each model on the lexical choice validation set.

5.2. Results of the Contrastive Task

The accuracy and inconsistency scores of these combinations of prompts are shown in Figure 1 on the anaphora task, for all 12 LLMs. The numbers are given in Table 2 for the five most interesting LLMs, on the development and test sets. The systems are ranked by decreasing accuracy of the best overall prompt, which is the one with simple system prompt and reasoning user prompt. The findings from experiments with the validation sets are confirmed on the test sets. Similarly, the scores for the lexical choice task are shown in Figure 2, and the scores for five LLMs on both subsets of the benchmark data are shown in Table 3.

To the best of our knowledge, the performance of previous systems is as follows. The creators of DiscEvalMT designed a system which encodes and decodes jointly the pairs of sentences, reaching ac-

Model	Development		Test	
	Acc.↑	Inc.↓	Acc.↑	Inc.↓
nllb-200	0.68	–	0.64	–
gpt-4o	0.84	0.14	0.90	0.07
	0.82	0.15	0.92	0.06
	0.80	0.22	0.88	0.12
	0.88	**0.01**	**0.97**	0.01
gpt-4	**0.88**	**0.01**	0.94	0.04
	0.86	0.04	0.95	0.01
	0.86	0.04	0.90	0.01
	0.78	0.05	0.88	0.00
gpt-4-turbo	0.86	0.04	0.91	0.00
	0.82	0.04	0.88	0.07
	0.84	0.05	0.88	0.09
	0.81	0.05	0.84	0.03
Phi-4 (14B)	0.78	0.16	0.90	0.01
	0.78	0.19	0.92	0.01
	0.79	0.04	0.89	0.04
	0.82	0.01	0.87	0.03
DeepSeek-R1 (32B)	0.69	0.04	0.81	0.02
	0.68	0.05	0.80	0.02
	0.64	0.10	0.68	0.02
	0.74	0.02	0.81	0.00

Table 2: Comparison of accuracy and inconsistency on development and test sets for the *anaphora task*. For each LLM, the four lines of scores are those of the prompts: (1) no system prompt and simple user prompt; (2) simple system and user prompts; (3) detailed system prompt and simple user prompt; (4) simple system prompt and step-by-step (reasoning) user prompt.

Model	Development		Test	
	Acc.↑	Inc.↓	Acc.↑	Inc.↓
nllb-200	0.68	–	0.54	–
gpt-4o	0.86	0.04	0.94	0.03
	0.88	0.04	0.94	0.04
	0.87	0.03	0.94	0.02
	0.96	0.02	**0.96**	0.01
gpt-4-turbo	0.87	0.01	0.92	0.03
	0.84	0.02	0.92	0.03
	0.84	0.03	0.92	0.04
	0.88	0.00	0.94	0.02
Phi-4 (14B)	0.83	0.00	0.89	0.01
	0.83	0.01	0.91	0.02
	0.84	0.07	0.90	0.03
	0.80	0.05	0.92	0.02
gpt-4	0.89	0.03	0.94	0.01
	0.90	0.04	0.92	0.02
	0.89	0.04	0.92	0.03
	0.92	0.02	0.91	0.02
DeepSeek-R1 (14B)	0.76	0.05	0.86	0.04
	0.76	0.01	0.88	0.03
	0.78	0.03	0.84	0.07
	0.79	0.03	0.78	0.01

Table 3: Comparison of accuracy and inconsistency on validation and test sets for the *lexical choice* task. For each LLM, the four lines correspond to the four configurations of prompts.

curacies of 0.72 and 0.57 respectively on anaphora and lexical cohesion (Bawden et al., 2018, Table 2). In a comparative evaluation study, Lopes et al. (2020, Table 7) found that the best performing approach was the concatenation of both sentences on both sides, as proposed by Tiedemann and Scherrer (2017), reaching 0.82 and 0.55 accuracies. The encoder-decoder model proposed by Pal et al. (2024, Table 4), which used a separate encoder for the context, reached 0.54 and 0.52 accuracy, while Zhang et al. (2022, Table 5) announced 0.64 accuracy on the anaphora task. A quality estimation approach with the COMET-QE metric (Vernikos et al., 2022) reached respectively 0.83 and 0.68 accuracy. Finally, our own use of NLLB-200 only reached, respectively, 0.66 and 0.61 accuracy (first line of Table 2).

On the anaphora task (Figure 1 and Table 2), the best results on the validation set came from GPT-4o with the reasoning prompt and from GPT-4 with the simplest prompt (no reasoning), both reaching 0.88 accuracy. However, GPT-4o with reasoning outperformed GPT-4 on the test set, reaching a nearly perfect accuracy of 0.97. For both systems, the presentation order of the alternatives had almost no influence on the answer, with an inconsistency score of only 0.1. GPT-4 also did well with the other prompts, except the reasoning one, scoring up to 0.86 with an inconsistency of 0.04, and GPT-4-turbo followed closely. GPT-3.5-turbo showed a larger variability, with inconsistency scores up to 0.45 on the validation set.

Among LLMs run locally, Phi-4 stood out with 0.82 accuracy with the reasoning prompt and very stable outputs (inconsistency of 0.01), making it the most reliable open-weight model. DeepSeek-R1 32B peaked at 0.74 accuracy also with a low inconsistency of 0.02, but smaller DeepSeek variants like the 14B and 8B models dropped to around 0.60 and were less consistent. Because Deepseek-R1 are reasoning models, the reasoning prompt did not bring significant benefits, apart from a small increase likely related to the instruction to "reason step-by-step" recommended by its authors. As expected from the usage instructions, a detailed system prompt was detrimental.

At the lower end, models like Llama 3.2 and Mistral struggled on the validation set, and were not tested any further. Llama 3.2 was totally unable to understand the reasoning prompt, reaching only 0.50 accuracy (random level) and an inconsistency of 0.98. Mistral scored between 0.54 and 0.59, with inconsistencies over 0.75 in several cases.

Overall, higher accuracy goes hand-in-hand with better consistency. The reasoning prompt was helpful for models that could handle it. For smaller or less instructed models, more complex prompts often led to worse outcomes or formatting issues.

On the lexical choice task (Figure 2 and Table 3) the trends are similar. The scores are even higher than for anaphora, with excellent consistency. The best performance came from GPT-4o with the reasoning-style prompt, reaching an accuracy of 0.96 and a very low inconsistency of 0.02. GPT-4 also performed strongly, with up to 0.92 accuracy, although its results varied slightly more across prompts. Again, models with strong reasoning capabilities, such as GPT-4o and GPT-4, consistently outperformed others, especially with the CoT prompt, and among open-source options, Phi-4 offered the best combination of performance and efficiency.

In terms of the length of response, which correlates with cost, DeepSeek-R1 models consistently generated long responses (over 600 tokens), regardless of the prompt, as they are trained to perform CoT reasoning. In contrast, most other models had output lengths that were more clearly shaped by the prompt. Llama 3.3 and Phi-4 produced some of the longest reasoning responses apart from DeepSeek-R1, with 279 and 227 tokens on average under the reasoning prompt. On the contrary, GPT-3.5-turbo, Mistral, and Llama 3.2 generated very short outputs, even with reasoning prompts, rarely exceeding 10 tokens. Therefore, given the higher cost of CoT responses, reasoning prompts are only worth using when they improve performance, as in the case of GPT-4o and Phi-4.

6. Translation Task

We now test the ability of LLMs to correctly translate a sentence when inter-sentential constraints – pronominal anaphora and lexical cohesion – are involved. Indeed, as observed by Post and Junczys-Dowmunt (2024) for encoder-decoder MT systems, scores on the contrastive task are not necessarily correlated with translation quality.

Using the same dataset, we give each LLM the two source sentences (EN) and the reference translation of the first sentence (FR). We obtain the translation of the second sentence from the LLM and score it against the reference translation present in the dataset using the four evaluation metrics presented in Section 3.2 above: BLEU, chrF, BERTScore, and COMET. As prompt engineering relied less on the development set than for the first task, we report below the results on the full set of 200 examples for each task.

6.1. Prompts: Reasoning or Not

Our goals are again to determine the best performance of LLMs on the two benchmarks, and to find if CoT reasoning improves performance over translation with no reasoning. The no-reasoning prompt is the same for both benchmarks, as shown in Appendix B.1. The system part contains general instructions for translation, while the user part simply provides the two EN sentences and the FR translation of the first one.

For the reasoning prompts (also in Appendix B.1), we make explicit in the system prompt several steps that guide the reasoning for each task. An additional instruction at the end requires the LLM to enclose the reasoning between XML-like tags, which leads to an easier to parse output, as well as similar or better performance.

6.2. Results of the Translation Task

On the anaphora benchmark, the translation scores with or without reasoning, and their differences (Δ), are shown in Table 4. The results show that the GPT family delivers the strongest performance, and the largest increase when reasoning is encouraged. The smaller "turbo" versions benefit most from reasoning in terms of BERTScore and COMET. Phi-4 exhibits the highest improvement on all metrics when reasoning is used. Together with Llama 3.3, they are the highest scoring open-weight models, though Llama 3.3 draws only modest benefits from reasoning. The DeepSeek-R1 models present a mixed picture: only the largest model (32B) improves its BLEU score, though not the other scores.

Models with weaker baselines, i.e., DeepSeek-R1 8B and 14B, Llama 3.1 and 3.2, and Mistral decrease across all metrics when prompted to reason. The scores of NLLB-200 place it midway, behind GPT, Llama 3.3 and Phi, but ahead of DeepSeek-R1 and Mistral. A possible explanation for the scores of DeepSeek-R1 is that, being a reasoning model, the additional reasoning instructions in the user prompt (beyond "please reason step by step") have a detrimental effect, together with the system prompt, especially for the smaller models.

On the lexical choice benchmark, the ranking of the LLMs is similar to the one on anaphora (see Table 5). Within the GPT family, reasoning improves GPT-4, GPT-4o, and GPT-3.5-turbo across most metrics. While GPT-4-turbo shows an unexpected small decline, GPT-4 seems to make a slightly more effective use of the reasoning instructions, despite GPT-4o's stronger baseline score. Among open-weight models, Phi-4 again improves on all metrics with the CoT prompt, while LLama 3.3 and DeepSeek-R1 32B improve only in BLEU and chrF, Mistral gains slightly, and the other models degrade, while NLLB-200 is close to the bottom of the ranking. Lexical choice may pose a more difficult challenge compared to pronominal anaphora, as acceptable translations of ambiguous words are more numerous. This may explain why improve-

Model	BLEU			chrF			BERTScore			COMET		
	w/o	w/	Δ	w/o	w/	Δ	w/o	w/	Δ	w/o	w/	Δ
nllb-200	37	–	–	60	–	–	.90	–	–	0.87	–	–
gpt-4	**49**	53	+3.78	**70**	**71**	+1.77	**.92**	**.93**	+0.0033	**.92**	**.92**	+0.0006
gpt-4o	**49**	**54**	+5.35	**70**	**71**	+1.96	**.92**	.92	+0.0016	.91	.91	+0.0017
gpt-4-turbo	45	49	+4.17	67	67	+0.90	.92	.90	-0.0173	.91	.90	-0.0084
gpt-3.5-turbo	44	49	+5.22	66	68	+2.15	.91	.91	+0.0017	.91	.90	-0.0080
Llama 3.3	44	47	+2.68	66	67	+1.08	.92	.91	-0.0046	.90	.90	-0.0043
Phi-4	43	49	**+5.58**	64	68	**+4.39**	.91	.92	**+0.0096**	.88	.91	**+0.0270**
DeepSeek-R1 32B	35	39	+4.34	59	59	-0.04	.89	.84	-0.0551	.87	.84	-0.0286
Llama 3.1	34	30	-3.78	58	54	-3.35	.89	.87	-0.0213	.86	.85	-0.0130
DeepSeek-R1 14B	34	33	-1.06	58	53	-4.91	.89	.79	-0.0927	.86	.79	-0.0647
Mistral	27	27	-0.03	51	50	-0.80	.86	.85	-0.0084	.82	.80	-0.0182
Llama 3.2	27	23	-3.48	51	47	-3.83	.87	.83	-0.0390	.82	.76	-0.0579
DeepSeek-R1 8B	24	21	-2.48	50	45	-4.75	.86	.77	-0.0908	.80	.77	-0.0298

Table 4: Translation quality scores on the *anaphora benchmark*: without reasoning (w/o), with reasoning (w/), and difference (Δ) between the latter and the former. Values for w/o and w/ are rounded to 2 digits, and Δ values to 4. Positive values of Δ indicate progress due to reasoning.

Model	BLEU			chrF			BERTScore			COMET		
	w/o	w/	Δ	w/o	w/	Δ	w/o	w/	Δ	w/o	w/	Δ
nllb-200	31	–	–	51	–	–	.86	–	–	0.77	–	–
gpt-4o	**54**	**54**	+0.31	**70**	**69**	-0.26	**.92**	**.92**	-0.0015	**.89**	**.88**	-0.0054
gpt-4	51	**54**	+2.58	67	**69**	+2.16	.91	**.92**	+0.0051	.86	.87	+0.0070
gpt-4-turbo	50	49	-0.31	66	65	-0.83	.91	.89	-0.0209	.87	.86	-0.0126
gpt-3.5-turbo	47	49	+2.11	65	66	+0.81	.91	.91	+0.0041	.86	.86	-0.0059
Llama3.3	46	47	+0.15	62	63	+0.90	.90	.90	-0.0018	.85	.84	-0.0030
Phi-4	43	44	+1.77	60	61	+1.74	.89	.90	+0.0064	.83	.84	+0.0021
DeepSeek-R1 32B	39	41	+2.07	56	58	+2.01	.88	.88	-0.0011	.82	.81	-0.0054
DeepSeek-R1 14B	38	33	-5.10	55	51	-3.99	.87	.85	-0.0226	.80	.77	-0.0330
Llama3.1	35	29	-5.71	52	49	-3.53	.87	.86	-0.0144	.79	.77	-0.0261
DeepSeek-R1 8B	31	30	-0.63	48	48	-0.28	.86	.83	-0.0227	.76	.75	-0.0115
Mistral	27	28	+1.17	45	45	+0.11	.84	.85	+0.0046	.74	.75	+0.0101
Llama3.2	25	22	-2.81	42	43	+0.63	.83	.82	-0.0121	.73	.70	-0.0256

Table 5: Translation quality scores on the *lexical choice benchmark*: without reasoning (w/o), with reasoning (w/), and difference (Δ) between the latter and the former. Values for w/o and w/ are rounded to 2 digits, and Δ values to 4. Positive values of Δ indicate progress due to reasoning.

ments through reasoning are slightly less consistent for the lexical choice task.

6.3. The "Wise Get Wiser" Effect

In both the anaphora and lexical choice experiments, we observe a surprising effect. Unlike many techniques which tend to improve weaker models but do not benefit the top-scoring ones, here reasoning is more beneficial to the LLMs which already score highest without it. Table 6 shows the Pearson correlations and the Spearman rank correlations between baseline scores without reasoning and the improvements obtained with reasoning (Δ), for both tasks. For anaphora, high coefficients for all metrics (0.59–0.81) confirm a consistent "wise get wiser" pattern: models that start from higher baselines tend to benefit more from reasoning. The effect is weaker for lexical choice, although correlations remain positive (0.21–0.52).

Moreover, we confirm that the variations of translation scores between the prompts without reasoning and those with reasoning are consistent across all metrics. This is important because scoring LLM translations uses MT metrics which correlate imperfectly with human preferences, unlike the accuracy metric of the first task. Figure 3, upper part, shows correlations among the Δ values of all systems, between all pairs of metrics. The rather large coefficients (Pearson: 0.58–0.91, Spearman: 0.66–0.92) indicate that models scoring well with respect to the others on one metric tend to score well on the other metrics as well. Similar values are observed for the lexical choice task, shown in Figure 3, lower part (Pearson: 0.71–0.90, Spearman: 0.71.–0.83).

7. Conclusion

This paper evaluated the capacity of several LLMs to pass two benchmarks for contextual MT, one targeting pronoun translation and the other one tar-

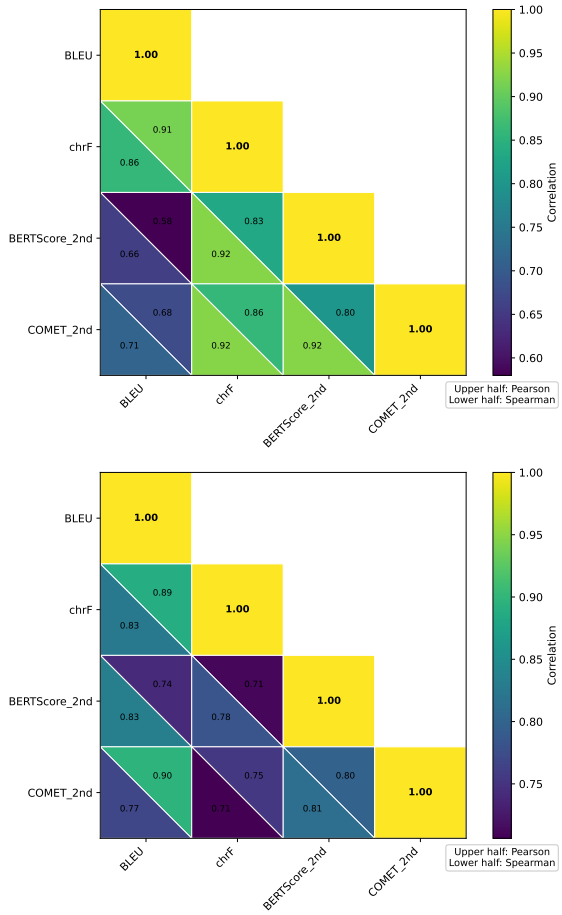

Figure 3: Correlations between the Δ values (reasoning minus no-reasoning) obtained with several MT metrics for 12 LLMs, on anaphora translation (upper part) and lexical choice translation (lower part). Each cell shows Pearson (upper triangle) and Spearman (lower triangle) correlations.

geting lexical coherence. We compared several types of prompts, and showed that CoT prompting elicits reasoning steps that lead to the best results. We found that the best models improve considerably the state of the art on the contrastive task, with accuracies slightly above 0.95 for both targeted phenomena. They also score highly on the four translation quality metrics, with COMET scores of 0.92 and 0.89. Moreover, we observed a "wise get wiser" effect, as the improvement brought by reasoning is positively correlated to the scores of the same LLMs without reasoning. In other words, the strongest models are also those that benefit the most from reasoning.

These results point to a possible future solution for improving translation with LLMs thanks to reasoning. The solution would need first to identify locations in documents where reasoning is likely to be beneficial, then generate the reasoning that makes translation choices explicit, separating it with markup from the actual translation. Either a generic CoT prompt could be used, or a specific one could be applied depending on the identified difficulties – as we did here with slightly different prompts for anaphora and lexical choice. Such a self-reflecting behavior could lend itself naturally to an agentic AI approach, in which a first-pass translation generated without reasoning could be improved by explicitly solving inter-sentential dependencies.

		BLEU	chrF	BERT Score	COMET
Anaphora	*P*	0.81	0.76	0.60	0.59
	S	0.68	0.72	0.64	0.74
Lexical choice	*P*	0.40	0.21	0.31	0.30
	S	0.52	0.27	0.30	0.25

Table 6: Pearson (*P*) and Spearman (*s*) correlations between the baseline scores without reasoning and the improvements due to reasoning (Δ). The large correlations show the "wise get wiser" effect.

8. Acknowledgments

We are grateful to the Swiss National Science Foundation for its support through the EXOMAT grant n. 228494, External Knowledge for Low-resource Machine Translation. We acknowledge the support of HES-SO through the grant n. AGP-140146. We would like to warmly thank the anonymous LREC reviewers for their insightful suggestions.

9. Bibliographical References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Rachel Bawden. 2018. *Going beyond the Sentence: Contextual Machine Translation of Dialogue*. Ph.D. thesis, Université Paris-Saclay.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large mul-

tilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process forimproved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Sheila Castilho and Rebecca Knowles. 2025. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? What about a GPT model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use ChatGPT. *NBER Working Paper 34255*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. *Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level*. MaTOS Deliverable D4-4.1.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, et al. 2025. R1-T1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella,

Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Daniil Larionov, Sotaro Takeshita, Ran Zhang, Yanran Chen, Christoph Leiter, Zhipin Wang, Christian Greisinger, and Steffen Eger. 2025. DeepSeek vs. o3-mini: How well can reasoning LLMs evaluate MT and summarization? *arXiv preprint arXiv:2504.08120*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models. *arXiv preprint arXiv:2503.10351*.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54(2).

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv 2207.04672*.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–483, Nusa Dua, Bali. Association for Computational Linguistics.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. *arXiv preprint arXiv:1901.09115*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2024. Evaluation and large-scale training for contextual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1125–1139, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 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)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. 2025. How well do large reasoning models translate? A comprehensive evaluation for multi-domain machine translation. *arXiv preprint arXiv:2505.19987*.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Appendix

A. Contrastive Task

A.1. Prompt Variants

This section presents the prompts used in the contrastive evaluation. Following common prompt engineering practice, we have experimented with various system prompts and user prompts. The prompts differ in their structures and instruction levels, while maintaining a consistent task format. Placeholders appear in curly braces in the prompt templates: `{context_en}`, `{source_sentence}`, `{context_fr}`, `{option1}`, `{option2}`. In the contrastive setting, two candidate translations are presented, and the model must output a constrained choice, either as "Choice: 1" or as "Choice: 2".

No system prompt, simple user prompt

User prompt

```
Here is a short text with two sentences
in English:
{context_en}
{source_sentence}
Here are two different translations into
French:
1. {context_fr}
{option1}
2. {context_fr}
{option2}
Which one is more correct?
Please answer 1 if the first one is more
correct or answer 2 if the second one is
more correct.  Do not add explanations.
MAKE SURE you only answer in the follow-
ing manner:  Choice:  (1 or 2)
```

Simple system and user prompts

System prompt

```
You are a language evaluation assistant.
Your task is to compare two French trans-
lations of an English text and decide
which is more correct.
When providing your answer, strictly fol-
low this format:  Choice:  (1 or 2)
Do NOT include any explanation or addi-
tional text.  Only output the specified
format.
```

User prompt

```
Here is a short text with two sentences
in English:
{context_en}
{source_sentence}
Here are two different translations into
French:
1. {context_fr}
{option1}
2. {context_fr}
{option2}
Which one is more correct?
Please answer 1 if the first one is more
correct or answer 2 if the second one is
more correct.
```

Detailed system prompt, simple user prompt

System prompt

```
You are a strict translation evaluation
assistant.  Your ONLY task is to deter-
mine which of two French translations of
an English text is more correct.
### Instructions:
- You will receive an English text con-
taining two sentences.
- You will be provided with two differ-
ent translations in French.
- Your task is to determine which trans-
lation is more accurate.
- You must respond in EXACTLY this for-
mat:  Choice:  (1 or 2)
- You MUST NOT provide any explanations,
thoughts, or additional text.
- Any deviation from the required format
is strictly prohibited.
- If you understand the instructions,
respond ONLY in the required format.
```

User prompt

```
Here is a short text with two sentences
in English:
{context_en}
{source_sentence}
Here are two different translations into
French:
1. {context_fr}
{option1}
```

```
2. {context_fr}
{option2}
Which one is more correct?
```

Simple system prompt, step-by-step (reasoning) user prompt for anaphora

System prompt

```
You are a helpful assistant. You have
to find which translation from English
to French is correct. Always keep your
answers short.
```

User prompt (reasoning)

```
Which one is more correct? Please an-
swer 1 if the first one is more correct
or answer 2 if the second one is more
correct.
Short text with two lines in English.
English line 1: {context_en}
English line 2: {source_sentence}
French translation of English line 1:
{context_fr}
French translation number 1 of line 2:
{option1}
French translation number 2 of line 2:
{option2}
Let's reason step by step to find the
correct French translation of line 2.
1. Find the difference between the two
translations of line 2.
2. Find the text in English line 2 which
is the cause of the translation differ-
ence at Step 1.
3. Find the text in English line 1 to
which the text found at Step 2 refers.
4. Find how the text from Step 3 is
translated in French translation of En-
glish line 1.
5. Find the correct word in French to
refer to this text.
6. Find which French translation in-
cludes this word.
Select 1 if French translation number 1
is more correct, or 2 if French transla-
tion number 2 is more correct.
When providing your answer, strictly
follow this format: Choice: (1 or 2)
```

Simple system prompt, step-by-step (reasoning) user prompt for lexical choice

System prompt

```
You are a helpful assistant. You have
to find which translation from English
to French is correct. Always keep your
answers short.
```

User prompt (reasoning)

```
Which one is more correct? Please an-
swer 1 if the first one is more correct
or answer 2 if the second one is more
correct.
Short text with two lines in English.
English line 1: {context_en}
English line 2: {source_sentence}
French translation of English line 1:
{context_fr}
French translation number 1 of line 2:
{option1}
French translation number 2 of line 2:
{option2}
Let's reason step by step to find the
correct French translation of line 2.
1. Find the difference between the two
translations of line 2.
2. Find the text in English line 2 which
is the cause of the difference at Step 1.
3. Find the text in English line 1 which
is identical to the text found at Step 2.
4. Find how the text from Step 3 is
translated in French translation of En-
glish line 1.
5. Find the correct word in French to
refer to this text.
6. Find which French translation in-
cludes this word.
Select 1 if French translation number 1
is more correct, or 2 if French transla-
tion number 2 is more correct.
When providing your answer, strictly
follow this format: Choice: (1 or 2)
```

A.2. Results: Anaphora

In Table 7 we provide the results on the contrastive anaphora benchmark for all LLMs and prompt types, in terms of accuracy and inconsistency scores. In Table 8 we provide the elapsed time in the conditions described in Section 4.

A.3. Results: Lexical Choice

In Table 9 we provide the results on the contrastive lexical benchmark for all LLMs and prompt types, in terms of accuracy and inconsistency scores. In Table 10 we provide the elapsed time in the conditions described in Section 4.

B. Translation Task

B.1. Prompt Variants

This section presents the full text of the prompts used in the translation task. Similarly to the contrastive prompts, each configuration has a *system prompt* and a *user prompt*, and the placeholders appear in curly braces: `{context_en}`, `{source_sentence}`, `{context_fr}`, `{option1}`, `{option2}`. In the gen-

Model	System and user prompts			
	None & simple	Both simple	Detailed & simple	Simple & step-by-step
gpt-4o	0.84 / 0.14	0.82 / 0.15	0.80 / 0.22	0.88 / 0.01
Phi-4 (14B)	0.78 / 0.16	0.78 / 0.19	0.79 / 0.04	0.82 / 0.01
gpt-4-turbo	0.86 / 0.04	0.82 / 0.04	0.84 / 0.05	0.81 / 0.05
gpt-4	0.88 / 0.01	0.86 / 0.04	0.86 / 0.04	0.78 / 0.05
DeepSeek-R1 (32B)	0.69 / 0.04	0.68 / 0.05	0.64 / 0.10	0.74 / 0.02
LLaMA 3.3 (70B)	0.74 / 0.27	0.74 / 0.27	0.74 / 0.20	0.69 / 0.04
DeepSeek-R1 (14B)	0.60 / 0.04	0.65 / 0.06	0.60 / 0.08	0.60 / 0.06
gpt-3.5-turbo	0.78 / 0.10	0.77 / 0.25	0.79 / 0.19	0.60 / 0.45
Mistral (7B)	0.54 / 0.78	0.50 / 1.00	0.62 / 0.42	0.59 / 0.27
LLaMA 3.1 (8B)	0.74 / 0.09	0.72 / 0.22	0.71 / 0.04	0.55 / 0.24
DeepSeek-R1 (8B)	0.48 / 0.09	0.49 / 0.14	0.48 / 0.02	0.54 / 0.05
LLaMA 3.2 (3B)	0.53 / 0.60	0.52 / 0.81	0.56 / 0.39	0.50 / 0.98

Table 7: Accuracy (↑) and inconsistency (↓) across all models and prompts on the **anaphora** validation set (separated by '/' in each cell). The systems are ranked by decreasing accuracy of the best overall prompt, which is the fourth one.

Model	System and user prompts			
	None & simple	Both simple	Detailed & simple	Simple & step-by-step
Phi-4 (14B)	0.36	0.39	0.36	2.13
DeepSeek-R1 (32B)	10.02	10.26	9.15	8.49
LLaMA 3.3 (70B)	3.03	2.64	2.46	62.43
DeepSeek-R1 (14B)	6.93	5.01	4.50	6.84
Mistral (7B)	0.30	0.27	0.27	0.39
LLaMA 3.1 (8B)	0.30	0.33	0.30	1.35
DeepSeek-R1 (8B)	4.11	4.41	4.62	3.42
LLaMA 3.2 (3B)	0.27	0.27	0.33	0.42
gpt-4o	0.228	0.258	0.234	0.768
gpt-4-turbo	0.294	0.276	0.252	0.894
gpt-4	0.798	0.960	1.308	3.060
gpt-3.5-turbo	0.192	0.240	0.198	0.204

Table 8: Mean elapsed time per prompt in seconds, across prompt configurations on the **anaphora** validation set. Open-source (Ollama) models listed first, followed by OpenAI models.

erative setting, the model must produce the French translation of the second sentence, with output format rules depending on the reasoning prompt used.

Prompt without reasoning – identical for anaphora and lexical choice

System prompt

```
You are a professional translator. Your
task is to translate short English texts
into French.

### Instructions:
- You will receive two English sentences:
a context sentence and a sentence to
translate.
- You will also be given the French
translation of the context sentence.
- Translate ONLY the second English sen-
tence into French.
- Return ONLY the French translation of
the second sentence.
- Do NOT include any explanation or addi-
tional text.
```

User prompt

```
Here is a short text with two sentences
in English:
{context_en}
{source_sentence}
Here is the French translation of the
first sentence:
{context_fr}
Please translate the second sentence
into French.
```

Structured reasoning (XML-style) – Anaphora

System prompt

```
You are a professional translator. Your
task is to translate short English texts
```

Model	System and user prompts			
	None & simple	Both simple	Detailed & simple	Simple & step-by-step
gpt-4o	0.86 / 0.04	0.88 / 0.04	0.87 / 0.03	0.96 / 0.02
gpt-4	0.89 / 0.03	0.90 / 0.04	0.89 / 0.04	0.92 / 0.02
gpt-4-turbo	0.87 / 0.01	0.84 / 0.02	0.84 / 0.03	0.88 / 0.00
LLaMA 3.3 (70B)	0.86 / 0.04	0.86 / 0.02	0.86 / 0.00	0.82 / 0.03
Phi-4 (14B)	0.83 / 0.00	0.83 / 0.01	0.84 / 0.07	0.80 / 0.05
gpt-3.5-turbo	0.80 / 0.05	0.83 / 0.05	0.84 / 0.04	0.80 / 0.08
DeepSeek-R1 (14B)	0.76 / 0.05	0.76 / 0.01	0.78 / 0.03	0.79 / 0.03
Mistral (7B)	0.58 / 0.69	0.50 / 1.00	0.64 / 0.46	0.76 / 0.19
DeepSeek-R1 (32B)	0.75 / 0.04	0.75 / 0.00	0.76 / 0.08	0.74 / 0.02
DeepSeek-R1 (8B)	0.67 / 0.07	0.60 / 0.11	0.62 / 0.12	0.66 / 0.02
LLaMA 3.1 (8B)	0.76 / 0.13	0.78 / 0.01	0.77 / 0.05	0.64 / 0.12
LLaMA 3.2 (3B)	0.66 / 0.32	0.58 / 0.57	0.68 / 0.24	0.52 / 0.90

Table 9: Accuracy (↑) and inconsistency (↓) across all models and prompts on the **contrastive lexical choice** validation set (separated by '/' in each cell). The systems are ranked by decreasing accuracy of the best overall prompt, which is the fourth one.

Model	System and user prompts			
	None & simple	Both simple	Detailed & simple	Simple & step-by-step
LLaMA 3.3 (70B)	2.49	2.79	2.16	52.53
Phi-4 (14B)	0.33	0.36	0.30	1.95
DeepSeek-R1 (14B)	4.53	3.51	3.21	4.47
Mistral (7B)	0.27	0.24	0.27	0.33
DeepSeek-R1 (32B)	8.55	7.92	6.27	7.29
DeepSeek-R1 (8B)	2.97	3.09	2.82	2.61
LLaMA 3.1 (8B)	0.30	0.27	0.30	1.35
LLaMA 3.2 (3B)	0.24	0.24	0.24	0.39
gpt-4o	0.408	0.498	0.444	1.788
gpt-4	0.504	0.480	0.480	3.432
gpt-4-turbo	0.552	0.564	0.576	1.404
gpt-3.5-turbo	0.348	0.348	0.360	0.396

Table 10: Mean elapsed time per prompt (seconds) across prompt configurations on the **lexical choice** validation set. Open-source (Ollama) models listed first, followed by OpenAI models.

```
into French.

### Instructions:
- You will receive two English sentences:
a context sentence and a sentence to
translate.
- You will also be given the French
translation of the first sentence.
- Translate the second English sentence
into French.
- To achieve this, think step by step
to resolve pronouns and references cor-
rectly using the context:
1. Identify pronouns/references in the
second sentence.
2. Find their referent in the first sen-
tence.
3. Check how that referent is translated
in the French context.
4. Choose the correct French pro-
noun/reference.
- Finally, output only this XML (valid
single root):
<result>
<reasoning></reasoning>
<answer></answer>
</result>
```

User prompt

```
Here is a short English passage:
{context_en}
{source_sentence}
Here is the French translation of the
first sentence:
{context_fr}
Please translate the second sentence
into French, following the instructions.
Output only the XML format specified.
```

Structured reasoning (XML-style) – Lexical Choice

System prompt

```
You are a professional translator.  Your
task is to translate short English texts
into French.

### Instructions:
- You will receive two English sentences:
a first sentence and a second sentence.
- You will also receive the French trans-
lation of the first sentence.
- Translate the second English sentence
into French.

To ensure lexical consistency:
1. Identify any key word or expression
in the second sentence that could be
translated in more than one way into
French.
2. Check whether the French translation
of the first sentence already provides a
preferred translation for that word or
expression,
and use the same choice if appropriate
for your translation of the second sen-
tence.

- Finally, output only this XML (valid
single root):
<result>
<reasoning></reasoning>
<answer></answer>
</result>
```

User prompt

```
Here is a short English text with two
sentences:
{context_en}
{source_sentence}
Here is the French translation of the
first sentence:
{context_fr}
Please translate the second sentence
into French, following the instructions
in the system prompt.
Output only the XML format specified.
```