

# Analyzing Context Contributions in LLM-based Machine Translation

Emmanouil Zaranis<sup>1,2</sup>, Nuno M. Guerreiro<sup>1,2,3,4</sup>, André F. T. Martins<sup>1,2,3</sup>  
<sup>1</sup>Instituto de Telecomunicações, <sup>2</sup>Instituto Superior Técnico, <sup>3</sup>Unbabel, <sup>4</sup>MICS  
emmanouil.zaranis@tecnico.ulisboa.pt

## Abstract

Large language models (LLMs) have achieved state-of-the-art performance in machine translation (MT) and demonstrated the ability to leverage in-context learning through few-shot examples. However, the mechanisms by which LLMs use different parts of the input context remain largely unexplored. In this work, we provide a comprehensive analysis of context utilization in MT, studying how LLMs use various context parts, such as few-shot examples and the source text, when generating translations. We highlight several key findings: (1) the source part of few-shot examples appears to contribute more than its corresponding targets, irrespective of translation direction; (2) finetuning LLMs with parallel data alters the contribution patterns of different context parts; and (3) there is a positional bias where earlier few-shot examples have higher contributions to the translated sequence. Finally, we demonstrate that inspecting anomalous context contributions can potentially uncover pathological translations, such as hallucinations. Our findings shed light on the internal workings of LLM-based MT which go beyond those known for standard encoder-decoder MT models.

## 1 Introduction

Large language models (LLMs) have reached state-of-the-art performance in machine translation (MT) and are making significant strides toward becoming the *de facto* solution for neural MT (Kocmi et al., 2023; Alves et al., 2024). Compared to the classical standard approach using encoder-decoder models (Bahdanau et al., 2016; Vaswani et al., 2017), LLMs are typically decoder-only models parameterized by billions of parameters. Remarkably, LLMs have demonstrated the ability to perform translation tasks without being explicitly trained for them, instead leveraging in-context learning (ICL) through demonstrations of the task (Zhang et al., 2022; Agrawal et al., 2023; Hendy et al.,

2023; Alves et al., 2023; Garcia et al., 2023). Yet, there is a gap in the literature on understanding the internal workings of LLM-based MT. Previous interpretability research on MT has been limited to traditional, specialized encoder-decoder models (Ding et al., 2017; Ferrando et al., 2022a,b; Voita et al., 2021; Sarti et al., 2024; Mohammed and Nicolae, 2024), and while substantial work has investigated ICL in other tasks, such as classification (Min et al., 2022; Lu et al., 2022; Yoo et al., 2022; Wang et al., 2023) and question answering (Liu et al., 2022; Liu et al., 2023; Si et al., 2023; Wei et al., 2023), the mechanisms by which LLMs leverage parts of context in MT remain largely unexplored.

In this work, we aim to fill this research gap by contributing towards a better understanding of how LLMs utilize different parts of the provided context (*e.g.*, few-shot examples, the source text, or previously generated target tokens) in MT. While previous work conducted on understanding the impact of context in MT largely focuses on performing modifications on the LLM input and measuring performance drop (Zhu et al., 2023; Raunak et al., 2023), we take instead an attribution-based approach (Ferrando et al., 2022a), tracking the input tokens’ relevance in all parts of the context—this allows us to estimate how different parts of context contribute to the generated translations, providing a more fine-grained analysis of context utilization.

We study several key aspects of context utilization in MT using general purpose LLaMA-2 models (Touvron et al., 2023) and TOWER models (Alves et al., 2024)—a suite of models specifically adapted for translation tasks. First, we investigate how different input parts contribute to the translated sequence. Next, we explore whether the provided few-shot examples contribute equally to the translated sequence. We also analyze if undergoing adaptation via continuous pretraining (Gupta et al., 2023; Çağatay Yıldız et al., 2024; Alves et al., 2024) on relevant multilingual and parallel

use this paper to prove the varying importance of input context positions

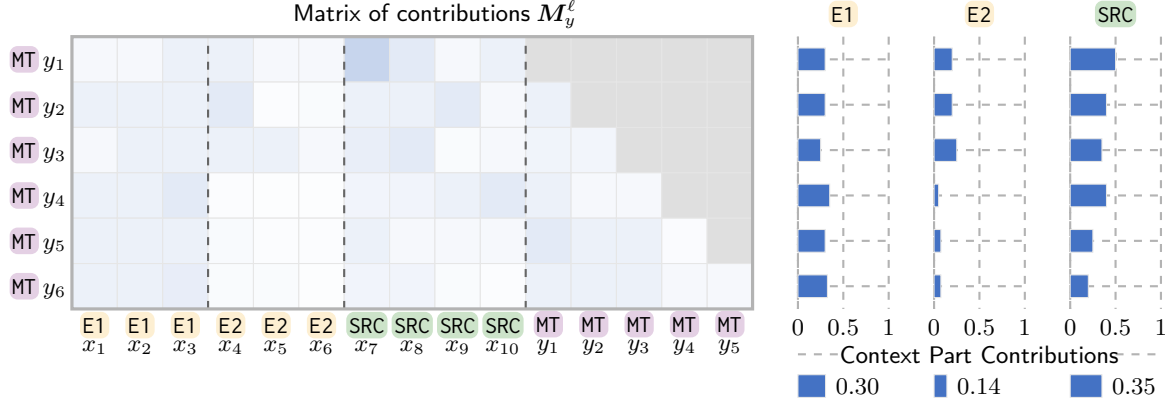


Figure 1: Illustration of *synthetic* part-level *total* contributions computation given 2 examples as context. From the token-to-token level contribution matrix  $M_y^\ell$ , we compute the total contribution of each input part to each generated token, by summing the corresponding token-level contributions. Subsequently, we compute the part-level total contribution of each input part to the translated sequence, by averaging over the generated tokens.

data leads to a change in these contribution patterns. Moreover, to further understand the translation dynamics, we examine how context contributions vary at different stages of the generation process. Finally, we also assess whether anomalous context contributions can uncover catastrophic translations, such as hallucinations (Dale et al., 2023a).

Our analysis reveals several key insights on context utilization by LLMs for translation, including:

- Irrespective of the translation direction, the source of each few-shot example contributes more than its corresponding target;
- The examined models exhibit a positional bias—earlier few-shot examples tend to have higher contributions to the translated sequence. Additionally, the bias is maintained across different generation stages;
- Training on task-specific data reduces the influence of few-shot examples and consequently shrinks the positional bias observed;
- Low source contributions can potentially uncover pathological translations.

We release all our code, and make our results available across all tested models and languages.<sup>1</sup>

## 2 Problem Formulation

In this section, we introduce ICL and describe how we employ the ALTI method (Ferrando et al., 2022a) to measure the contribution of each input *part* in the context to the translated sequence.

<sup>1</sup>[https://github.com/deep-spin/interp\\_llm](https://github.com/deep-spin/interp_llm)

### 2.1 In-Context Learning (ICL)

ICL is a paradigm where LLMs "learn" to solve new tasks at inference time by being provided with a few task demonstrations as part of the input prompt, without requiring any updates to their parameters or fine-tuning (Brown et al., 2020; Agrawal et al., 2023; Hendy et al., 2023). More broadly, for MT, few-shot examples can also be used for inference time adaptation, *e.g.* to different domains, terminology, or other elements of translation, guiding the model to produce outputs that are more suitable for the given context (Alves et al., 2023; Aycock and Bawden, 2024).

### 2.2 ALTI for autoregressive language models

For our analysis, we choose the ALTI (Aggregation of Layer-Wise Token-to-Token Interactions) method (Ferrando et al., 2022a) for its simplicity and proven success in various applications. ALTI has been successfully employed for detecting hallucinations in MT (Dale et al., 2023b; Guerreiro et al., 2023), identifying toxicity in multilingual text (Team et al., 2022; Costa-jussà et al., 2023), and explaining information flows in LLMs (Ferrando and Voita, 2024; Tufanov et al., 2024).

ALTI is an input attribution method that quantifies the mixing of information in the transformer architecture (Vaswani et al., 2017). It follows the modeling approach proposed by Abnar and Zuidema (2020), where the information flow in the model is simplified as a directed acyclic graph, with nodes representing token representations and edges representing the influence of each input token representation on the output token representation (for

each layer of the transformer). ALTI proposes using token contributions instead of raw attention weights, and computes the amount of information flowing from one node to another in different layers by summing over the different paths connecting both nodes, where each path is the result of the multiplication of every edge in the path. Formally, given an input sequence of length  $S$  and an output sequence of length  $T$ , we compute a token-to-token contribution matrix  $C^\ell \in \mathbb{R}^{(S+T) \times (S+T)}$ , where  $\ell$  is the  $\ell$ -th layer of the model.<sup>2</sup> The element  $c_{i,j}^\ell$  of the matrix represents the contribution of the  $j$ -th input token at layer  $\ell - 1$  to the  $i$ -th output token at layer  $\ell$ . By multiplying the layer-wise coefficient matrices,  $M^\ell = C^\ell \cdot C^{\ell-1} \dots C^1$  we can describe representations of intermediate layers (and final layer) as a linear combination of the model input tokens—an example of a contribution matrix is shown in Figure 1.<sup>3</sup> This matrix can be used to interpret the model’s behavior and study how different parts of the input influence generated outputs. For more details, see Ferrando et al. (2022a).

### 2.3 Part-level contributions

To quantify the contribution of each input part to the translated sequence, we perform a two-step aggregation process, illustrated in Figure 1. First, we compute the total contribution of each part to each generated token by summing the corresponding token-level contributions within each part (right hand-side of Figure 1). Then, we average the part-to-token contributions across the generated tokens to compute the contributions of each context part to the entire translated sequence. Similarly to (Ferrando et al., 2022a; Dale et al., 2023a,b; Guerreiro et al., 2023), these part-level contributions are used for the analysis in the following sections.<sup>4</sup>

## 3 Experimental Setup

We provide an overview of the models and datasets used throughout our study, as well as important considerations on how we prompt the models.

**Models.** We experiment with two families of models: the general-purpose LLAMA-2 7B base model (Touvron et al., 2023), and the state-of-the-art TOWER 7B base model, which is a continued

pretrained checkpoint of LLAMA-2 7B on a mixture of monolingual and parallel data (Alves et al., 2024). We also experiment with TOWERINSTRUCT 7B, which is obtained via finetuning TOWER on a set of instructions for translation-related tasks.<sup>5</sup>

**Datasets.** We conduct our study on the publicly available WMT22 test sets, examining English to German (en-de), German to English (de-en), English to Russian (en-ru) and Russian to English (ru-en) language pairs.<sup>6</sup>

**Few-shot setting and prompt selection.** We conduct our analysis under a 5-shot setting, using the few-shot examples provided by Hendy et al. 2023, which were selected to be high-quality examples and relevant—according to embedding similarity—to the source text. We make sure that the examples in the context are shuffled and not sorted by relevance to the source.<sup>7</sup> We use the prompt templates suggested in Zhang et al. 2023. Additional details are provided in Appendix A.1.

**Filtering.** Due to the high GPU memory requirements of the attribution method when applied to a 7B parameter model, we had to filter samples with large context length. We provide more details about the filtering process in Appendix A.2.

## 4 How Do Different Context Parts Contribute to the Translated Sequence?

In this section, we conduct a top-level analysis by measuring and comparing the contributions of different input parts to the generated translation.

### 4.1 Analysis setup

To investigate the contribution of different prompt parts to the translated sequence, we first divide the context into the following parts: source and target side of each few-shot example, source text, and target prefix. Then, we follow the approach described in Section 2.3 and obtain part-level contributions that are used for analysis.

<sup>5</sup>We use the following HuggingFace checkpoints: LLAMA-2 (meta-llama/Llama-2-7b-hf), TOWER (Unbabel/TowerBase-7B-v0.1), and TOWERINSTRUCT (Unbabel/TowerInstruct-7B-v0.2).

<sup>6</sup>German is the second most frequent language in LLAMA-2 (Touvron et al., 2023), just behind English, while Russian accounts for approximately 0.13% of the training data.

<sup>7</sup>We include experiments with a different shuffling seed in Appendix B—trends in results are similar to those reported in the main text.

<sup>2</sup>Note that this matrix is causal masked.

<sup>3</sup>For simplicity, we will consider  $M_y^\ell$  as the matrix containing the last  $T$  rows of  $M^\ell$ —these rows contain the contributions of the input parts to the output tokens.

<sup>4</sup>We follow previous work and analyze the last-layer contributions.

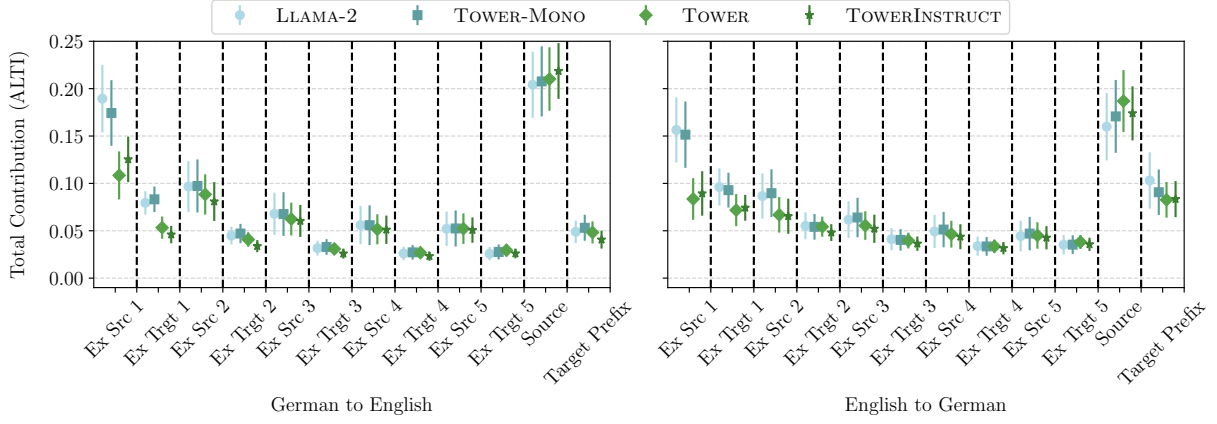


Figure 2: Illustration of context’s part-level contributions to the translated sequence, for all the examined models.

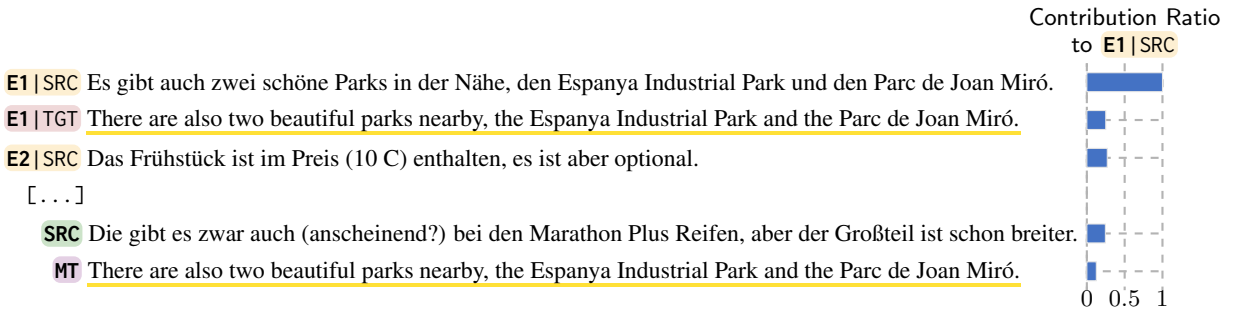


Figure 3: Example of anomalous source contributions for TOWER which hallucinates, copying information from the first example. We show contribution ratios to **E1 | SRC**—1 being the contribution of **E1 | SRC**.

## 4.2 Results

In Figure 2, we show, for all the examined models, the total contribution of each context part to the translated sequence for German to English and English to German language pairs.

**The source of each few-shot example consistently contributes more than its corresponding target.** For each of the examined models, we notice that the source of each provided example is more influential than the corresponding target for generating the translation. This finding is consistent across language pairs. Aligning with findings in classical encoder-decoder MT models (Ferrando et al., 2022a; Guerreiro et al., 2023), where it was found that models tend to have higher source text contribution when translating into English than out of English, we find that the source contribution, both at the example and test source level, is higher for German to English than in English to German.

**Training on parallel data reduces the impact of the provided examples on the translated sequence.** We observe that the contributions of few-shot examples, particularly the first examples, are

much greater for LLAMA-2 than for both TOWER models. One hypothesis is that the continued pre-training with parallel data on TOWER makes it rely less on the examples since it is not required to “learn” the task “on-the-fly”. This leads to an interesting question: *what if we replace the parallel data and instead only use monolingual data for multiple languages?* To investigate this, we examine the TOWER-MONO model.<sup>8</sup> Interestingly, we find that TOWER-MONO behaves much more similarly to LLAMA-2 than TOWER. This suggests that continual pretraining with task-specific data may lead the model to rely less on examples to perform the task. Exploring how to train dedicated models to be better guided by in-context examples is an interesting direction for future work.

**Close inspection of context contributions can uncover anomalous translations.** Previous works in neural MT have connected trends in context

<sup>8</sup>TOWER-MONO was trained following the same training procedure as TOWER (Alves et al., 2024). The only difference to the former is that, instead of using 20B tokens of text split in 2/3 monolingual data and 1/3 parallel data, it was trained with 20B tokens of monolingual data.



contributions, particularly low source contributions, to pathological translations such as hallucinations (Ferrando et al., 2022a; Dale et al., 2023b; Guerreiro et al., 2023). Through close inspection of our analyzed samples, we indeed find a series of pathological translations. Figure 3 presents one such example—here, the source contribution is particularly low, representing only about 25% of the contribution of the first example; interestingly, the generated translation is, in fact, an exact copy of the translation from that first example. We provide additional examples in Appendix B.3. We will return to these and other salient cases in Section 6 to examine how contributions evolve for such cases during the generation process.

**A clear positional trend emerges in few-shot example contributions.** Figure 2 shows a remarkable “stair-like” trend in the contribution of few-shot examples to the translated sequence. On average, the influence of each example appears to be strongly correlated with its position in the context, with earlier examples exhibiting higher contributions than later ones. This suggests there may be a positional bias in how the models leverage the provided examples during the translation process.

## 5 Examining Positional Bias over the Provided Few-shot Examples

Motivated by the findings from the previous section, we now closely inspect properties of the positional bias in few-shot example contributions.

### 5.1 Are examples that occur early in the context more influential than later ones?

Here we perform a sample-level analysis to obtain a better understanding of the relationship between examples’ contributions and their respective position. Specifically, we aim to explore whether there is a systematic and monotonic relationship between the order of few-shot examples and their contributions.

#### 5.1.1 Analysis setup

We examine whether the contributions of the first  $K$  few-shot examples monotonically dominate the remaining  $N - K$  examples, where  $N$  is the total number of examples used in the context. In other words, for each sample, we check if the contributions of the first  $K$  examples are sorted in descending order and if they are strictly higher than the contributions of the remaining  $N - K$  exam-

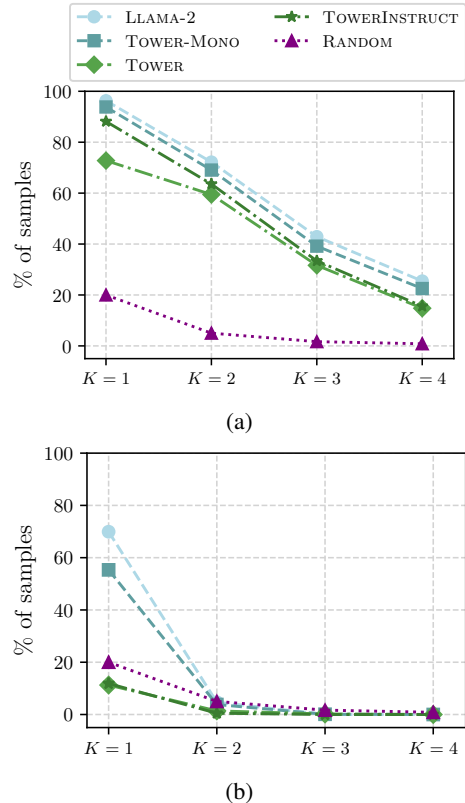


Figure 4: Proportion of de-en samples that follow positional bias, for different values of  $K$ , in the (a) original and (b) replace-last-ex settings.

ples.<sup>9</sup> We consider different values of  $K$  to represent different types of positional bias. For instance, when  $K = 1$ , the first few-shot example attains the highest level of contribution. When  $K = 4$ , the few-shot examples exhibit globally monotonic contributions, indicating a strong positional bias across all examples. Examples for each bias type are provided in Appendix C.

To quantify the prevalence of each type of positional bias, we measure the proportion of samples that satisfy the aforementioned condition for each value of  $K$ . We then compare these proportions to the probability, under a permutation of the examples drawn uniformly at random (denoted as RANDOM), of the first  $K$  few-shot examples monotonically dominating the remaining  $N - K$  examples, which is given as  $p = (N - K)!/N!$ .

#### 5.1.2 Results

We show results for German to English translation in Figure 4a.<sup>10</sup>

<sup>9</sup>We do not require the contributions of the remaining  $N - K$  examples to be monotonically sorted.

<sup>10</sup>We include results for the rest language pairs examined in Appendix C—trends are largely similar.

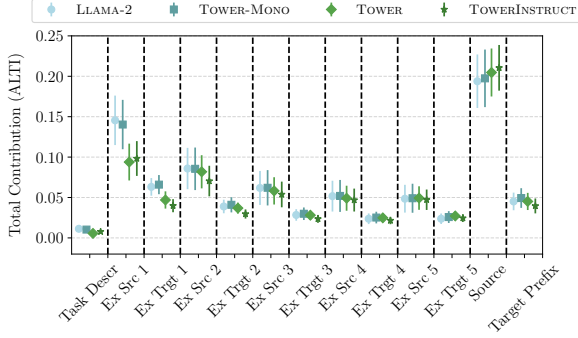


Figure 5: Illustration of context’s part-level contributions, when the task description is added. Translation direction: *German to English*

**Positional bias is prevalent and follows a monotonic pattern.** Our analysis reveals that positional bias is significantly more common than the RANDOM baseline for all values of  $K$ , suggesting that it is a prevalent phenomenon in the examined models. Additionally, we observe a monotonic relationship: the bias is more frequent for the first few examples than for later ones. This implies that the influence of positional bias gradually decreases as we move further down the context.

**The bias is particularly stark for the first few-shot examples.** All models tend to assign higher contribution to the first example, with this bias being more prevalent for models not trained on parallel data. For these models, over 95% of the analyzed samples exhibit the highest contribution for the first example.<sup>11</sup> Models trained with parallel data, either through continued pretraining or additional finetuning, show a slight decrease in the first-example bias, but it remains significant compared to the RANDOM baseline.

The observed positional bias raises an important question: *are contributions merely a function of position or are they connected to content of the context parts?* We will conduct two additional experiments in the next section to inspect this phenomenon closer.

## 5.2 How strong is the positional bias?

We now turn to a more detailed investigation of the positional trend we found in the results above. Specifically, we investigate how the introduction of other context parts and the relevance of the examples interact with the trend.

<sup>11</sup>We remark again that the examples in the context are shuffled and not sorted by relevance to the source.

### 5.2.1 Is it all about position?

First, we examine the impact of adding a task description before the examples.<sup>12</sup> If the bias is solely position-dependent, we might expect the task description to receive higher contribution due to its placement at the beginning of the context. This analysis will help us understand whether the positional bias is influenced by the nature of the content or if it is strictly position-based.

**Task description receives minimal contribution despite its position.** The results of our first experiment, shown in Figure 5, reveal that, despite appearing at the beginning of the input text, the task description receives significantly lower contribution compared to the examples and other parts of the context. This suggests that the positional bias is not merely a function of absolute position, but may rather depend on the nature of the content. Interestingly, even though a new part of context was added, the positional bias over the examples—“stair-like” trend in the contributions—is still present.

### 5.2.2 Can relevance to the test example break the bias?

We now investigate whether an overwhelmingly relevant example can break the positional bias, even when it appears later in the context.

To test this, we create an artificial setup—replace-last-ex—where a copy of the test example (source and translation) is placed as the last example in the context. Intuitively, if the model is shown a source text along with its corresponding translation in the context, the most straightforward approach would be to copy the translation. As such, we expect the model to assign higher contribution to this last example, overriding the positional bias.

**The bias is shrunk significantly.** Figure 4b shows that this intervention significantly reduces the positional bias, particularly for the TOWER and TOWERINSTRUCT models. In contrast, for models not trained on parallel data, the first example still contributes more than all other examples—even when a copy is present in the context—way more frequently than random chance. Interestingly, the bias is almost entirely broken for all other example positions. These findings suggest that while relevant content can indeed shrink the bias, the first ex-

<sup>12</sup>We can assume the “task description” as an additional part of the context. We use the following description template: *Translate the following text from [SRC\_LANG] to [TGT\_LANG]v<sub>i</sub>.*

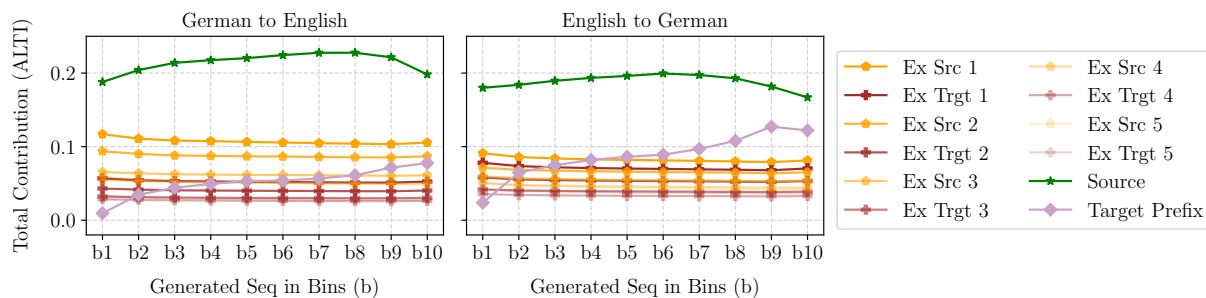


Figure 6: Illustration of how context contributions evolve across different generation stages for the TOWER model. Each generated bin accounts for 10% of the generated sequence.

amples influence the translation generation beyond simply “solving the task.” They likely provide additional cues, such as the language pair and expected output format, that shape the model’s behavior.

## 6 How Do Context Contributions Evolve during the Generation Process?

In the previous sections, we examined which parts of the provided context have the greatest influence on the translated sequence. We now shift our focus to explore how these context contributions evolve across different stages of the generation process.

### 6.1 Analysis setup

To investigate this, we divide the generated sequence into 10 bins of equal length and compute the total contribution of each context part to each bin. We then average these contributions across samples to obtain a comprehensive view of how the influence of different context parts changes as the translation progresses.

**Results.** In Figure 6, we present the average total contribution of each individual part to each generated bin, for the TOWER models.

**Relative ranking of context parts’ contributions remains stable throughout generation.** We observe that the relative ranking of contributions from different context parts is largely preserved throughout the generation process. Specifically, the source text consistently exhibits the highest contribution across all bins, followed by the few-shot examples in descending order of their position—this reinforces the notion of positional bias. The only exception to this pattern is the target prefix, which attains higher contribution as it grows in length. This is expected: with a longer prefix, the model increasingly relies on the previously generated tokens to inform its predictions. Moreover, we also

find a decrease in the source contribution at the last stage of generation, suggesting that the model relies less on the source when generating the final tokens. Interestingly, both these observations align with findings in traditional neural MT models, which have shown similar patterns in the relative contributions of source and target information during the generation process (Voita et al., 2021).

**Translation direction impacts the evolution of context contributions.** While the overall ranking of context part contributions remains similar, we observe notable differences when translating into or out of English. As noted earlier in Section 4, the source contribution is higher when translating into English (de-en) compared to when translating out of English (en-de). Interestingly, in de-en translation, the source of each example also consistently contributes more than its corresponding target, resulting in a “stacked” appearance of source contributions—the contribution from any example’s source is bigger than that of any example’s target text. In contrast, en-de translation exhibits an alternating contribution ranking, with the source and target of each example interleaved (e.g., src example 1 > tgt example 1 > src example 2 > tgt example 2, and so on). Moreover, we also observe that the target prefix contribution grows much more steeply in en-de than in de-en, suggesting that when translating a non-English text, the model relies more heavily on the context (examples and source) throughout the generation process.

**Highlighting the importance of source-part contributions in anomalous cases.** Building on our findings from Section 4, which showed that close inspection of context contributions can potentially uncover anomalous translations, we further analyze such cases in terms of how context contributions evolve during the generation process. We compare

<b>E1   SRC</b>	Es gibt auch zwei schöne Parks in der Nähe, den Espanya Industrial Park und den Parc de Joan Miró.
<b>E1   TGT</b>	There are also two beautiful parks nearby, the Espanya Industrial Park and the Parc de Joan Miró.
<b>E2   SRC</b>	Das Frühstück ist im Preis (10 €) enthalten, es ist aber optional.
<b>E2   TGT</b>	Breakfast is included in the price (10 €), but it is optional.
<b>E3   SRC</b>	Es gibt auch kostenlose Internet 24/7 und WiFi in allen Zimmern.
<b>E3   TGT</b>	There is also free internet 24/7 and wifi in all rooms.
<b>E4   SRC</b>	Bisher gibt es noch keine Bewertungen für S-Plus Company!
<b>E4   TGT</b>	There are no reviews for S-Plus Company yet!
<b>E5   SRC</b>	Die Größe der Wohnung ist 15 m2, es ist klein, aber sehr gemütlich.
<b>E5   TGT</b>	The size of the apartment is 15 m2, it's small but very cosy.
<b>SRC</b>	Die gibt es zwar auch (anscheinend?) bei den MarathonPlus Reifen, aber der Großteil ist schon breiter.
LLAMA-2 ✓	
<b>MT</b>	There are also (apparently?) at Marathon Plus Tyres, but the majority is wider.
TOWER ✗	
<b>MT</b>	There are also two beautiful parks nearby, the Espanya Industrial Park and the Parc de Joan Miró.

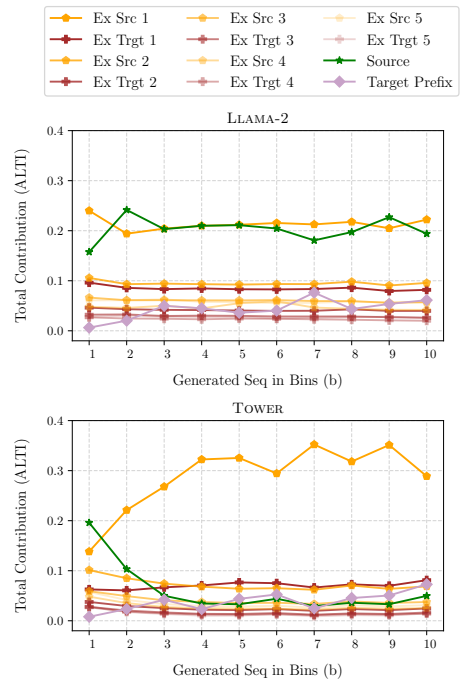


Table 1: Illustration of an example exhibiting anomalous source contributions for TOWER — which hallucinates, followed by LLAMA-2’s contributions, which performs normally.

the behavior of LLAMA-2 and TOWER models using the example presented in Table 1 (the same presented in Section 4). For LLAMA-2, which generates a correct translation, the context contribution trends align with the average case for German to English translation (see Figure 19 in Appendix D.1). In contrast, TOWER, which produces an incorrect translation by copying the first example, exhibits anomalous contribution trends (compared to Figure 6). Specifically, we observe a steeply increasing contribution from the first example, while the source contribution decreases significantly, highlighting the copying behavior. Additional salient cases are discussed in Appendix D.2.<sup>13</sup> Crucially, we find that in such cases, *source contributions*—both at the example and test source levels—can *potentially* indicate *pathological translations* and also provide insights into the factors driving the generation.

Language Pair	Model	AUROC
en-ru	LLAMA-2	52.3
de-en	TOWER	97.3
en-ru	TOWER	88.7

Table 2: AUROC of low source contribution scores.

<sup>13</sup>Here, we not only provide examples of other hallucinations, but also of other correct translations for which the context contributions follow interesting non-typical patterns.

**Low source contributions are, in some cases, predictive of hallucinations.** Our previous observations may potentially align well with previous neural MT research linking pathological translations to low source contributions (Ferrando et al., 2022a; Dale et al., 2023b; Guerreiro et al., 2023). Note again that classical encoder-decoder MT models and large language models (LLMs) are distinct in terms of the parts of context they often support: in classical encoder-decoder NMT models, the "context" for generation typically comprises only the source sentence and previously generated tokens; LLMs, however, often maintain a much broader context, potentially including various other relevant information. This distinction means that low source contribution in LLMs may not be so predictive of pathological translations, as the model might be drawing from other relevant contextual information. To explore this further, we conduct a quantitative analysis to assess the extent to which low-source contribution can be associated with hallucinations. Initially, for each model and language pair combination, we identify instances of "fully-detached" hallucinations by annotating the generated translations using the LLAMA-3-70B-INSTRUCT model (Dubey et al., 2024), following the exact approach outlined by Benkirane et al.,



2024.<sup>14</sup> For each model-language pair combination for which we observed a reasonable number<sup>15</sup> of "fully-detached" hallucinations, we report the AUROC of the low source contribution score in Table 2. Our findings suggest that while for TOWER low source contributions are particularly associated with hallucinations, it is not the case for LLAMA2. Upon closer inspection, we find that the low source contribution is particularly predictive of hallucinations that come in the form of exact copies of the provided few-shot examples<sup>16</sup>. Investigating these trends further, not only in machine translation but also in other tasks where context is relevant, is an interesting direction for future research.

## 7 Conclusion

We have comprehensively studied context contributions in LLM-based MT using the general purpose LLAMA-2 and translation-specialized TOWER models, exploring a broad range of key aspects, including investigating how different parts of context contribute to generated translations, and how these contributions evolve during the generation process.

Our findings reveal a strong positional bias, where earlier few-shot examples in the context have higher contributions to the translated sequence, both at the sentence level and across different generation stages. Interestingly, our experiments show that this bias is shrunk by continuous pretraining on task-specific data. Moreover, we reveal that the source part of each few-shot example has higher contribution compared to its corresponding target, irrespective of the translation direction. Finally, we stress the importance of source-part contributions by demonstrating that anomalous contributions can potentially uncover pathological translations, such as hallucinations. We believe our work not only provides insights into the internal workings of LLM-based MT, but also draws important connections to classical encoder-decoder NMT models.

To support future research on this topic, we are open-sourcing our code and releasing all data used in our analysis.

---

<sup>14</sup>In this paper, the authors show that this LLM can achieve performance comparable or even better than previously proposed detectors.

<sup>15</sup>We provide further quantitative results on the number of detected hallucinations in Appendix D.3.

<sup>16</sup>TOWER's pathological translations are usually copies of the few-shot examples, while this is not the case for LLAMA2.

## Limitations

While our study provides a valuable insight of how context is utilized by LLMs in MT, there are a few limitations that should be acknowledged.

Firstly, due to limitations in terms of computational resources paired with the fact that the ALTI method employed in our study can be computationally intensive, we restricted our analysis to 7B parameter models. This constraint raises the question of whether our findings still hold true when larger LLMs are considered, making it a potential direction for future studying.

Secondly, it should be noted that we focused exclusively on LLAMA-based models, particularly aiming on analyzing the TOWER-family of models, which are specifically oriented for MT. This selection enabled us to study how continued pretraining and finetuning on task-specific data impacts context utilization. However, this decision makes it so that it is still unclear whether our findings generalize to other LLM families.

Despite these limitations, we believe our study can lead to a better understanding of the dynamics of context utilization in LLM-based MT, providing key insights that can motivate future work on the field and inspire other research directions.

## Ethical Considerations & Potential Risks

Utilizing LLMs for MT might raise potential risks that should be pointed out, particularly regarding pathological translations and the ethical usage of contextual data.

Firstly, one of the critical risks which arises when using LLMs for MT is the phenomenon of pathological translations, such as hallucinations. As our study reveals, anomalous context contributions can potentially indicate these pathological translations, especially when low reliance on the source text is noticed. Despite the potential of detecting these pathological translations, their occurrence remains an important concern, as misinterpretations and incorrect translations might lead to significant consequences in specific domains such as healthcare, law etc. Thus ensuring that LLMs provide reliable translations is crucial.

Secondly, the reliance of LLMs in specific parts of the context when translating, introduces ethical considerations that should be taken into account regarding the choice of some context parts, such as the few-shot examples. The provided context might contain biases and misleading or inappropriate con-

tent and as a result this might be propagated into the generated translations. Our research can significantly contribute to mitigate this risk by identifying which parts of the provided context are responsible for propagating biases or inappropriate content to the translated sequence.

To conclude, addressing these risks and ethical considerations is important to foster a better usage of these systems and prevent potential harms.

## Acknowledgements

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

## References

- Samira Abnar and Willem Zuidema. 2020. [Quantifying attention flow in transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.
- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. [In-context examples selection for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.
- Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. [Steering large language models for machine translation with finetuning and in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. [Tower: An Open Multilingual Large Language Model for Translation-Related Tasks](#). *arXiv e-prints*, arXiv:2402.17733.
- Seth Aycock and Rachel Bawden. 2024. [Topic-guided example selection for domain adaptation in LLM-based machine translation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate](#). *Preprint*, arXiv:1409.0473.
- Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. [Machine translation hallucination detection for low and high resource languages using large language models](#). *Preprint*, arXiv:2407.16470.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. [Toxicity in multilingual machine translation at scale](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.
- David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. [Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.
- David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023b. [HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.
- Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. [Visualizing and understanding neural machine translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev,

Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Alonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Milon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shao-liang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenxin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng

Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymier, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Sweet, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damla, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pa-

- van Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratan-chandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiao-jian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022a. *Towards opening the black box of neural machine translation: Source and target interpretations of the transformer*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022b. *Measuring the mixing of contextual information in the transformer*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. *Explaining how transformers use context to build predictions*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.
- Javier Ferrando and Elena Voita. 2024. *Information flow routes: Automatically interpreting language models at scale*. *Arxiv*.
- Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. *The unreasonable effectiveness of few-shot learning for machine translation*. *Preprint*, arXiv:2302.01398.
- Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. *Hallucinations in Large Multilingual Translation Models*. *arXiv e-prints*, arXiv:2303.16104.
- Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. *Continual pre-training of large language models: How to (re)warm your model?* *Preprint*, arXiv:2308.04014.
- 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*.
- 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, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. *Findings of the 2022 conference on machine translation (WMT22)*. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. *What makes good in-context examples for GPT-3?* In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. *Lost in the Middle: How Language Models Use Long Contexts*. *arXiv e-prints*, arXiv:2307.03172.



- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Wafaa Mohammed and Vlad Niculae. 2024. [On measuring context utilization in document-level MT systems](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian’s, Malta. 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.
- Vikas Raunak, Arul Menezes, and Hany Awadalla. 2023. [Dissecting in-context learning of translations in GPT-3](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 866–872, Singapore. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. [Quantifying the plausibility of context reliance in neural machine translation](#). In *The Twelfth International Conference on Learning Representations*.
- Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. [Measuring inductive biases of in-context learning with underspecified demonstrations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11289–11310, Toronto, Canada. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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](#). *Preprint*, arXiv:2207.04672.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *arXiv e-prints*, arXiv:2307.09288.
- Igor Tufanov, Karen Hambardzumyan, Javier Ferrando, and Elena Voita. 2024. [Lm transparency tool: Interactive tool for analyzing transformer language models](#). *Arxiv*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2021. [Analyzing the source and target contributions to predictions in neural machine translation](#). In *Proceedings*

of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1126–1140, Online. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. [Label words are anchors: An information flow perspective for understanding in-context learning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. [Larger language models do in-context learning differently](#). *arXiv e-prints*, arXiv:2303.03846.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. [Ground-truth labels matter: A deeper look into input-label demonstrations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. 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, ICML’23*. JMLR.org.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [Opt: Open pre-trained transformer language models](#). *Preprint*, arXiv:2205.01068.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. [Multilingual machine translation with large language models: Empirical results and analysis](#). *ArXiv*, abs/2304.04675.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermiş. 2024. [Investigating continual pretraining in large language models: Insights and implications](#). *Preprint*, arXiv:2402.17400.

## A Further Details on Experimental Setup

### A.1 Few-shot setting & Prompt selection

We conduct our experiments using the few-shot examples provided by [Hendy et al. 2023](#), which were selected to be of high-quality and relevant to the source.

Following prior work ([Zhang et al., 2023](#)), we use the in-context template illustrated in Table 3.

```
SRC_LANG: E1 | SRC
TGT_LANG: E1 | TGT
SRC_LANG: E2 | SRC
TGT_LANG: E2 | TGT
[ . . . ]
SRC_LANG: SRC
TGT_LANG:
```

Table 3: Prompt template for few-shot inference.

### A.2 Filtering details

Due to our resource constraints, coupled with the high GPU memory requirements of the attribution method when applied to a 7B parameter model, we had to filter samples with large context length. More specifically, we exclude samples exceeding 400 tokens, when considering the concatenation of the input prompt with the generated sequence. We additionally filter out the samples for which the generated sequence does not exceed the length of 10 tokens.<sup>17</sup> We report the sizes of the sets—over 1000 samples for each language pair—examined in our analysis in Table 4.

Language Pair	Sample Size
De-En	1021
Ru-En	1017
En-De	1174
En-Ru	1107

Table 4: Sample sizes for each language pair considered in our analysis.

### A.3 Evaluation Details

We evaluate the models used in our work on all language directions examined to ensure high translation quality. We report BLEU ([Papineni et al., 2002](#)), COMET-22 ([Rei et al., 2022a](#)), and COMETKiwi ([Rei et al., 2022b](#)) in Table 5.

### A.4 Inference

We used greedy decoding at inference time, setting 300 tokens as the maximum length for the generated sequence.

<sup>17</sup>In our analysis in Section 6, we separate the generated sequences into 10 bins.

### A.5 Hardware specifications

All our experiments were conducted using 3 NVIDIA RTX A6000 GPUs.

### A.6 Discussion on artifacts

The data used for analysis in this paper was initially released for the WMT22 General MT task (Kocmi et al., 2022) and can be freely used for research purposes. All translation demonstrations (few-shot examples) used in our paper were released in (Hendy et al., 2023) under a MIT license.

Our code was developed on top of original ALTI repositories (Ferrando et al., 2022a, 2023), which have been released under Apache-2.0 License.

## B Top-level Analysis

In the top-level analysis conducted in Section 4, we examined the contributions of individual parts of the context to the translated sequence and highlighted several findings. In addition, we provide results for the Russian to English and English to Russian language pairs (§ B.1). As supplementary material, we include an additional experiment (§ B.2) to enhance the validity of our findings, and we also present examples exhibiting anomalous part-level contributions (§ B.3) for completeness.

### B.1 Context’s part-level contributions for additional language pairs

In Figure 7, we show, for all the examined models, the total contribution of each context part to the translated sequence for Russian to English and English to Russian language pairs. We observe that results are largely similar with those presented in the main text for the German to English and English to German language pairs.

### B.2 Additional experiment by reshuffling the order of few-shot examples

To ensure our findings hold against any potential, yet highly unlikely, content-related bias stemming from the position of the few-shot examples, we conduct a supplementary experiment. Put simply, we reshuffle the order of the few-shot examples for each sample and repeat the analysis. We report the results in Figures 8 and 9 for German and Russian languages respectively. The top-level part-level contributions remain largely consistent with those presented in the main text. This result underscores the validity of the findings presented in Section 4.

### B.3 Examples with anomalous part-level contributions

In Figures 10 and 11, we include some additional cases where the models hallucinate by copying one of the provided few-shot examples. We observe that in all cases the models exhibit anomalous contributions and particularly the contribution of the source is minimal. We also closely inspect similar cases in Appendix D.2, where we analyze the context dynamics across the generation stages and we discuss our findings.

## C Positional Bias Analysis

### C.1 Details on analysis setup and examples of positional bias types

In the analysis conducted in Section 5.1, we assess the prevalence and the extent of the positional bias observed. Particularly, we examine whether the contributions of the first  $K$  few-shot examples monotonically dominate the remaining  $N - K$  examples. We consider different values of  $K$  to represent the different types of positional bias. For instance, when  $K = 1$ , the first few-shot example attains the highest level of contribution. In the case where  $K = 2$ , the first two examples exhibit sorted contributions in a descending order and the remaining three have lower contributions than the first two, but they are not necessarily sorted in a descending order. Similarly, in the case where  $K = 3$ , the first three few-shot examples exhibit sorted contributions in a descending order and the remaining two have lower contributions than the first three, but they are not necessarily sorted in a descending order. Finally, when  $K = 4$ , the few-shot examples exhibit globally monotonic contributions, indicating a strong positional bias across all examples. We visually illustrate examples of the aforementioned cases in Figure 12.

### C.2 Additional plots

**Is it all about position?** In Figures 13, 14 and 15 we show the context’s part-level contributions, when the task description is added for the English to German, English to Russian and Russian to English translation directions respectively. We notice that in all translation directions the task description receives significantly lower contribution compared to the examples and other parts of the context, suggesting that the positional bias is not merely a function of absolute position.

	De-En			En-De		
	BLEU	COMET-22	COMETKiwi	BLEU	COMET-22	COMETKiwi
LLAMA-2	28.42	82.25	78.82	21.12	78.79	74.95
TOWER-MONO	28.19	82.45	78.90	23.42	80.99	77.88
TOWER	30.19	83.22	79.60	29.39	84.40	81.58
TOWERINSTRUCT	35.24	85.72	81.43	42.66	88.11	83.11

	Ru-En			En-Ru		
	BLEU	COMET-22	COMETKiwi	BLEU	COMET-22	COMETKiwi
LLAMA-2	32.99	82.53	78.84	20.03	80.78	76.80
TOWER-MONO	33.47	83.04	79.16	23.19	83.26	79.31
TOWER	37.78	83.84	79.79	28.33	86.10	82.03
TOWERINSTRUCT	44.48	86.53	81.51	40.02	89.72	83.41

Table 5: Translation performance of each examined model on the WMT22 test set.

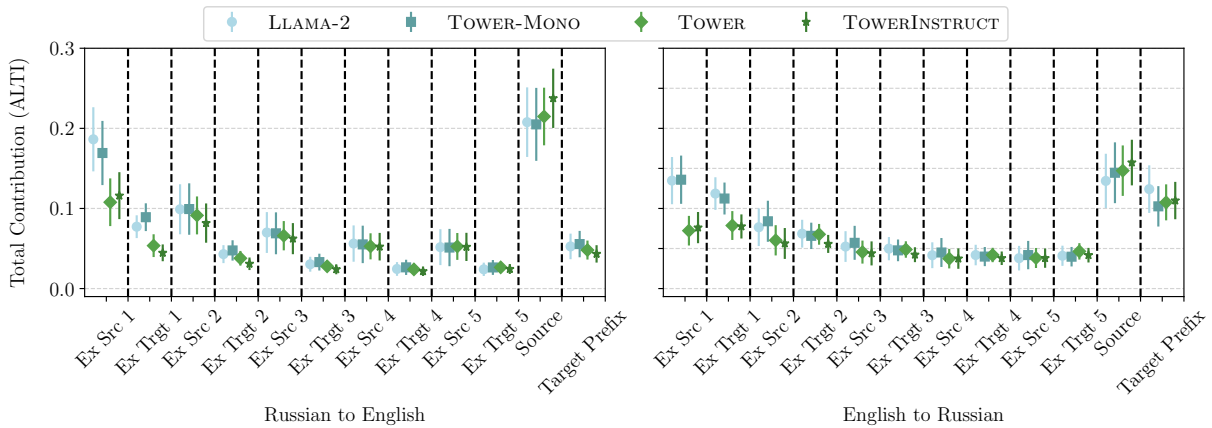


Figure 7: Illustration of context’s part-level contributions to the translated sequence, for all the examined models.

**Can relevance to the test example break the bias?** In Figures 16a and 16b, we present the proportion of en-de samples that follow positional bias, for different values of  $K$ , in the original and replace-last-example settings respectively. We additionally provide the corresponding results for the Russian to English and English to Russian translation directions in Figures 17 and 18 respectively. In all settings examined, we observe that results are largely similar with those presented in Sections 5.1 and 5.2.

## D Context Contributions across Generation Stages

In Section 6, we explored how context contributions evolve across different stages of the generation process for the TOWER model. In the following part, we include additional plots examining how context contributions evolve across the generation process for the rest of the models and language

pairs examined. We additionally show examples of anomalous context contributions and other salient cases and we discuss the results.

### D.1 Additional plots

In Figure 19, we present how context contributions evolve across different generation stages for LLAMA-2, TOWER-MONO and TOWERINSTRUCT models, for the de-en and en-de translation directions. For completeness, we provide in Figures 20 and 21 the corresponding plots for the ru-en and en-ru language pairs respectively.

### D.2 Examples of anomalous context contributions and other salient cases

In Section 6, we highlighted the importance of anomalous source-part contributions as indicators of pathological translations. Here, we include more such examples as well as instances of other salient cases.



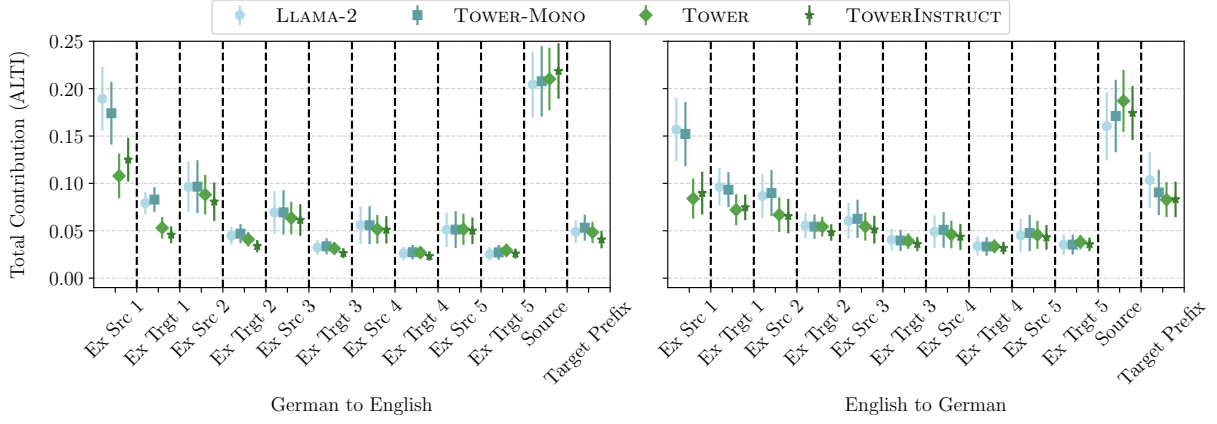


Figure 8: Illustration of context’s part-level contributions to the translated sequence, when reshuffling the order of provided few-shot examples.

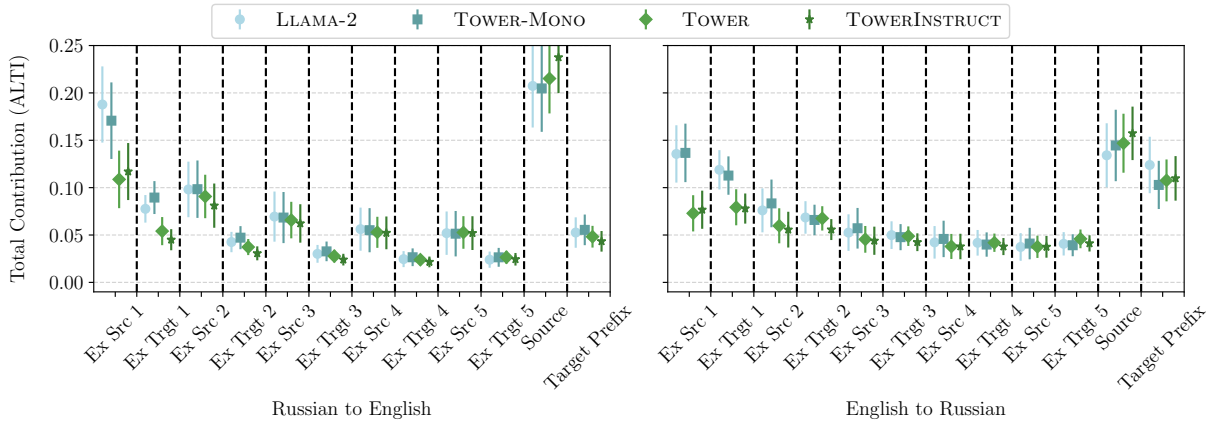


Figure 9: Illustration of context’s part-level contributions to the translated sequence, when reshuffling the order of provided few-shot examples.

In Tables 7, 8 and 9, we present 3 examples where one of the examined models hallucinates, exhibiting anomalous contributions. The example shown in Table 7 is particularly interesting, as both models in the beginning of the translation process exhibit low source contributions — compared to the source-part contribution of the first example — indicating that they primarily rely on the first example. However, as the translation progresses, the source contributions of the examined models follow completely opposite trends. TOWER exhibits extremely anomalous contributions — a steeply increasing contribution from the source-part of the first example and a decreasing one from the source — producing in this way a hallucination, by copying the first example. In contrast, LLAMA-2 produces a correct translation, with its contributions following the average case trends for German to English translation. Importantly, in all the provided examples, the models that produce a correct translation exhibit contribution trends that align with the aver-

age case trends we presented for German to English translation (see Figures 6 and 19 for TOWER and LLAMA-2 respectively).

Let’s now turn to some other salient cases. In particular, we now turn to examples where the models do not produce any pathological translations (see Tables 10 and 11). Note that the models exhibit low source contributions in the early steps of the translation process (compared to the contributions of the few-shot examples) indicating a greater influence from the few-shot examples that are semantically similar. Then, as the translation progresses, they exhibit increased source contributions being very similar with the average case trends for German to English translation (see Figures 6 and 19 for TOWER and LLAMA-2 respectively), indicating the reliance on the source to produce a correct translation.

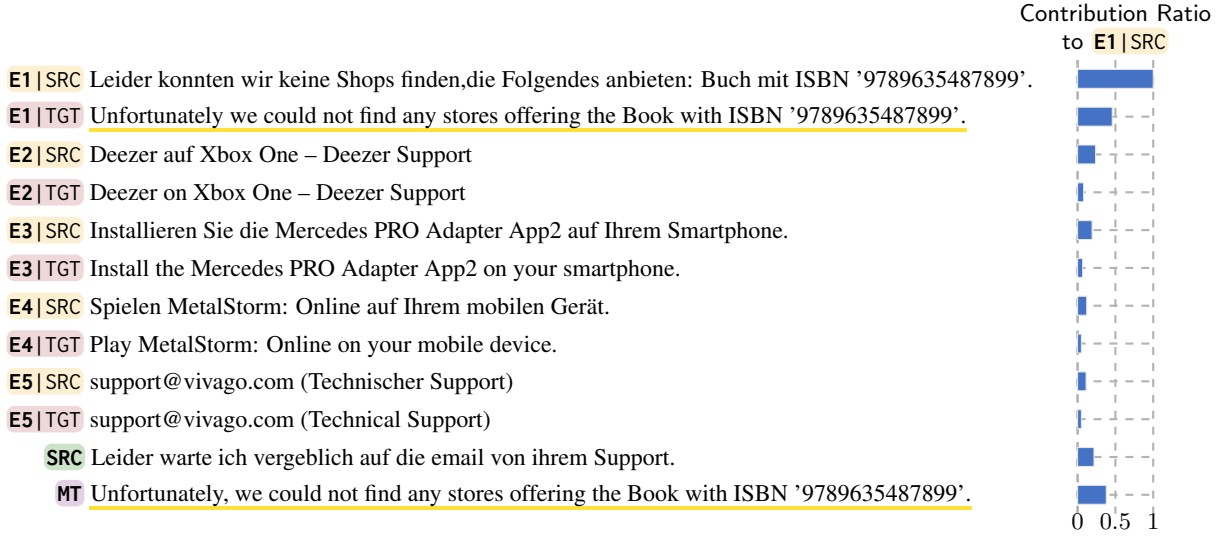


Figure 10: Example of anomalous source contributions for TOWER which hallucinates, copying information from the first example. We show contribution ratios to E1|SRC—1 being the contribution of E1|SRC.

Language Pair	Model	# of hall.
En-De	LLAMA-2	3
En-De	TOWER-MONO	4
En-De	TOWER	1
En-De	TOWERINSTRUCT	1
De-En	LLAMA-2	2
De-En	TOWER-MONO	2
De-En	TOWER	11
De-En	TOWERINSTRUCT	0
En-Ru	LLAMA-2	23
En-Ru	TOWER-MONO	4
En-Ru	TOWER	10
En-Ru	TOWERINSTRUCT	1
Ru-En	LLAMA-2	1
Ru-En	TOWER-MONO	5
Ru-En	TOWER	2
Ru-En	TOWERINSTRUCT	1

Table 6: Number of fully detached hallucination cases by language pair and model.

### D.3 Details of Quantitative Analysis

In Section 6, we examined whether anomalous context contributions can serve as indicators of hallucinations. Specifically we focused on how low source contributions, by conducting a quantitative analysis to assess the extent to which low-source contributions can be associated with "fully-detached" hallucinations. In this section, we provide further details regarding the annotation process.

For each model and language pair combination, we identify instances of "fully-detached" hallucinations by annotating the generated translations using the LLAMA-3-70B-INSTRUCT model (Dubey et al., 2024), following the exact approach outlined

by Benkirane et al., 2024.<sup>18</sup> Specifically, each instance is annotated into one of four categories: "No hallucination", "Small hallucination", "Partial hallucination", and "Full hallucination". Only instances classified as "Full hallucination" are considered "fully-detached" hallucinations in our analysis. We report the number of full hallucinations for each of model and language pair combination in Table 6.

### E AI Assistants

We have used Github Copilot<sup>19</sup> during development of our research work.

<sup>18</sup>We used the "Severity Ranking Prompt 2" as this was shown to be the optimal prompt for LLAMA-3-70B-INSTRUCT.

<sup>19</sup><https://github.com/features/copilot>

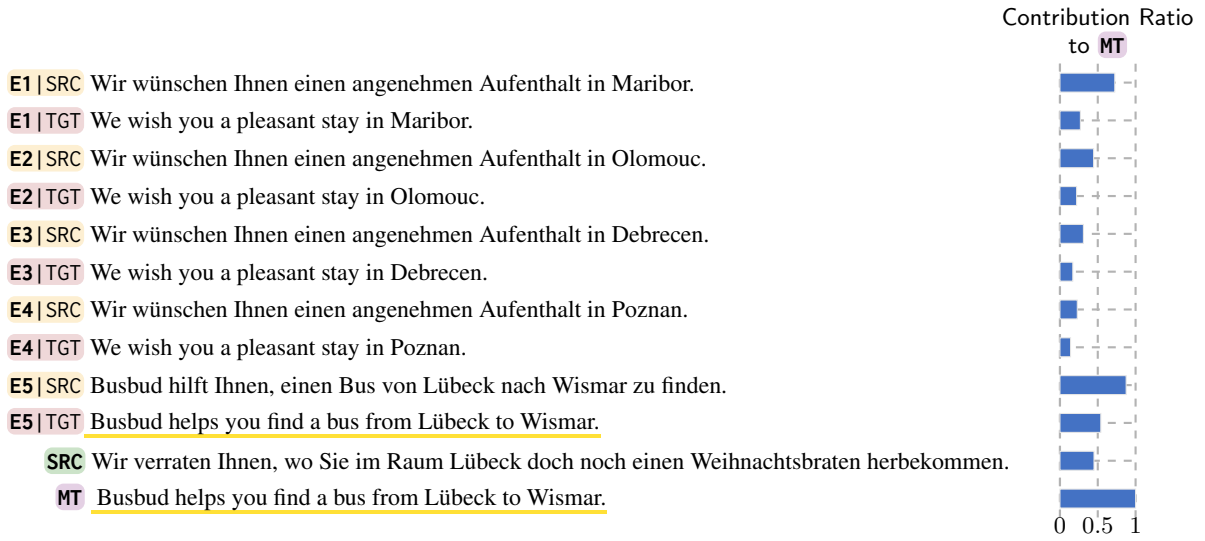


Figure 11: Example of anomalous source contributions for TOWER which hallucinates, copying information from the last example. We show contribution ratios to **MT**—1 being the contribution of **MT**.

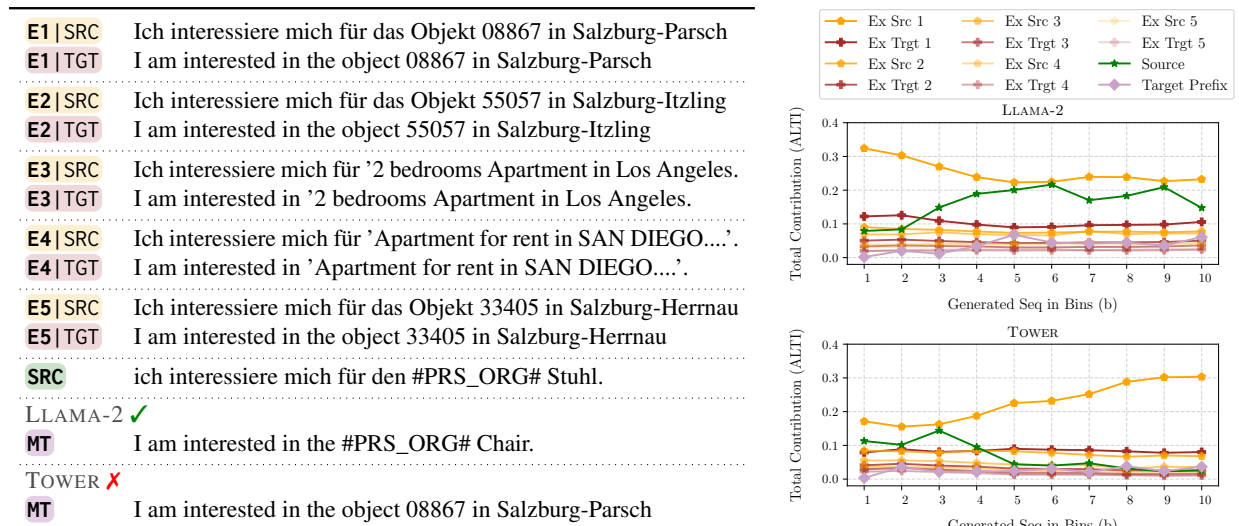
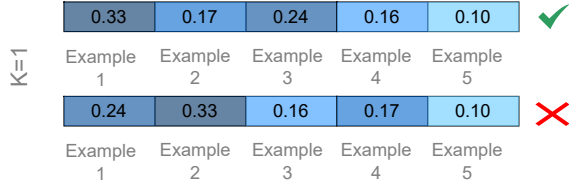
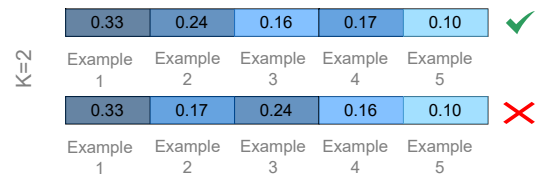


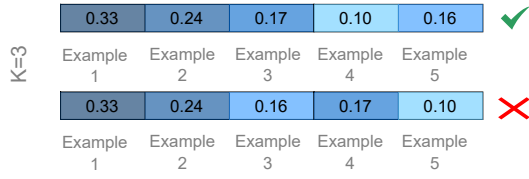
Table 7: Illustration of an example exhibiting anomalous source contributions for TOWER — which hallucinates, followed by LLAMA-2’s contributions, which performs normally.



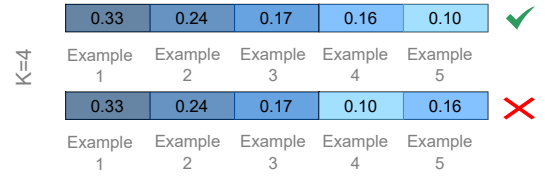
(a) The top sample follows the examined positional bias ( $K = 1$ ) as the first example attains the highest contribution. The bottom sample does not follow the bias, as the second example has greater contribution than the first.



(b) The top sample follows the examined positional bias ( $K = 2$ ) as the first two examples monotonically dominate the remaining three and the last three have lower contributions than the first two. Note that the last three examples do not necessarily exhibit sorted contributions in decreasing order. The bottom sample does not follow the bias, as the third example has greater contribution than the second.



(c) The top sample follows the examined positional bias ( $K = 3$ ) as the first three examples monotonically dominate the remaining two and the last two have lower contributions than the first three. Note that the last two examples do not necessarily exhibit sorted contributions in decreasing order. The bottom sample does not follow the bias, as the fourth example has greater contribution than the third.



(d) The top sample follows the examined positional bias ( $K = 4$ ) as the contributions of all the examples are sorted in decreasing order. The bottom sample does not follow the bias, as the fourth example breaks the monotonicity.

Figure 12: For each of the examined positional bias types we illustrate 2 examples. One that follows the examined type of positional bias and one that does not. We note that the demonstrated examples are provided for purely illustrative purposes and do not depict any real data.

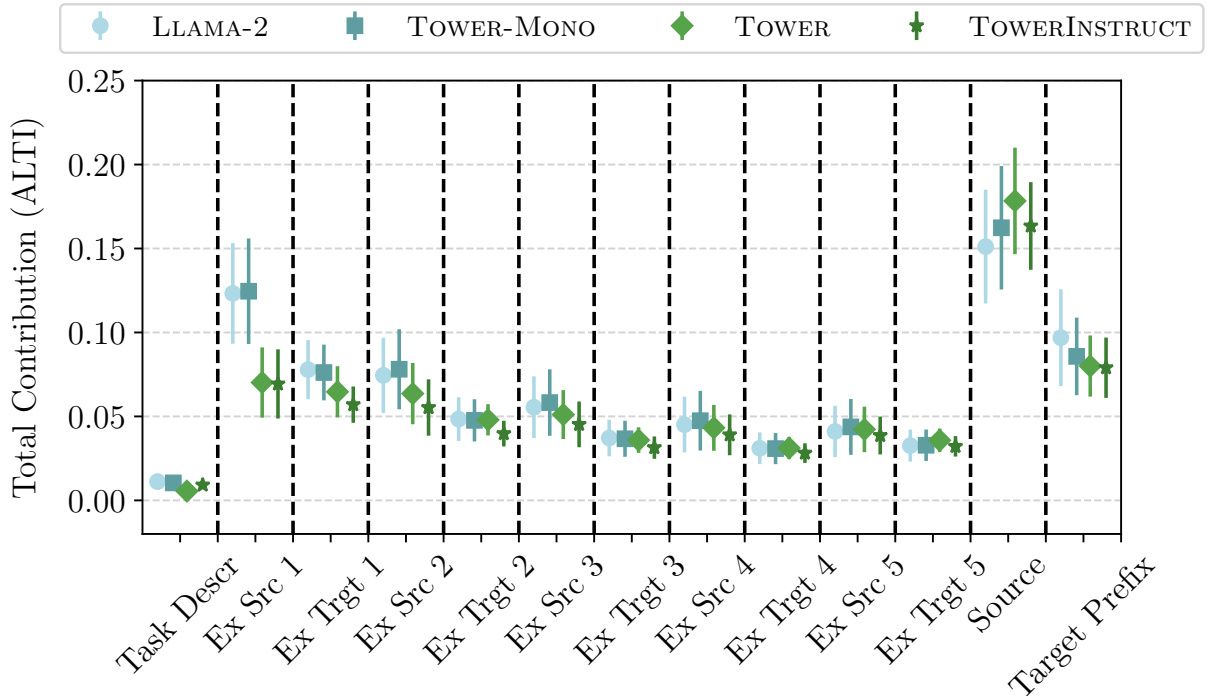


Figure 13: Illustration of context's part-level contributions, when the task description is added. Translation direction: *English to German*



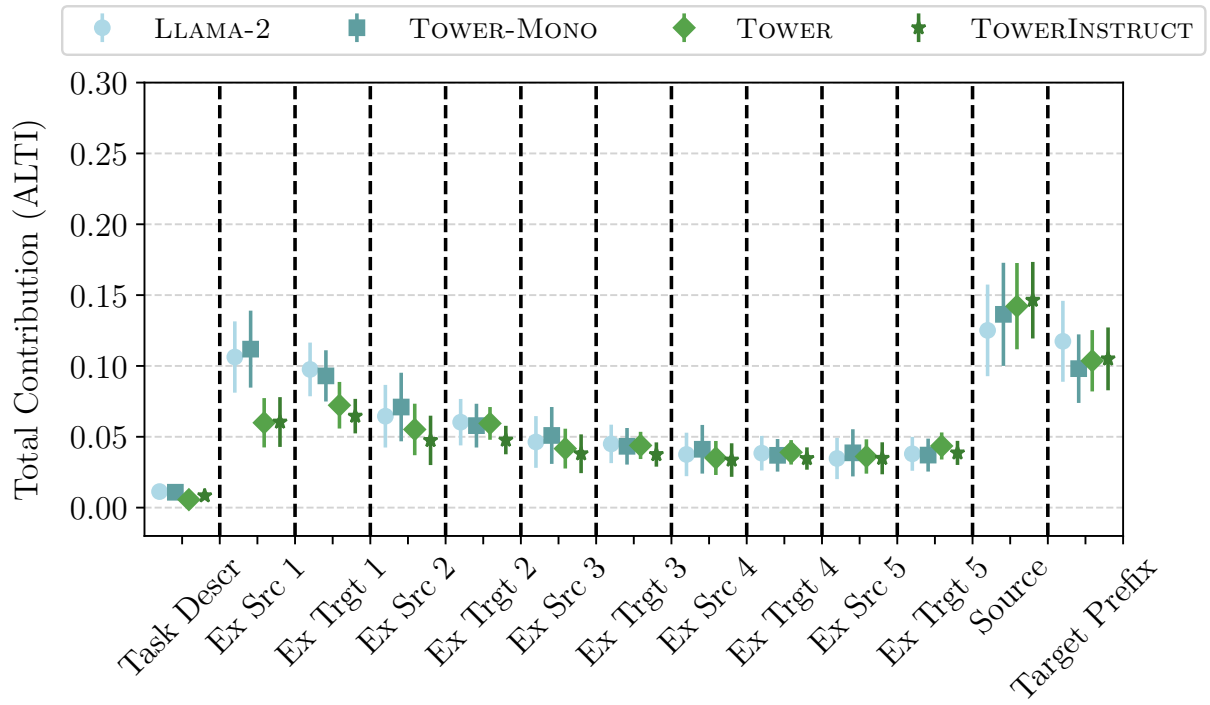


Figure 14: Illustration of context’s part-level contributions, when the task description is added. Translation direction: *English to Russian*

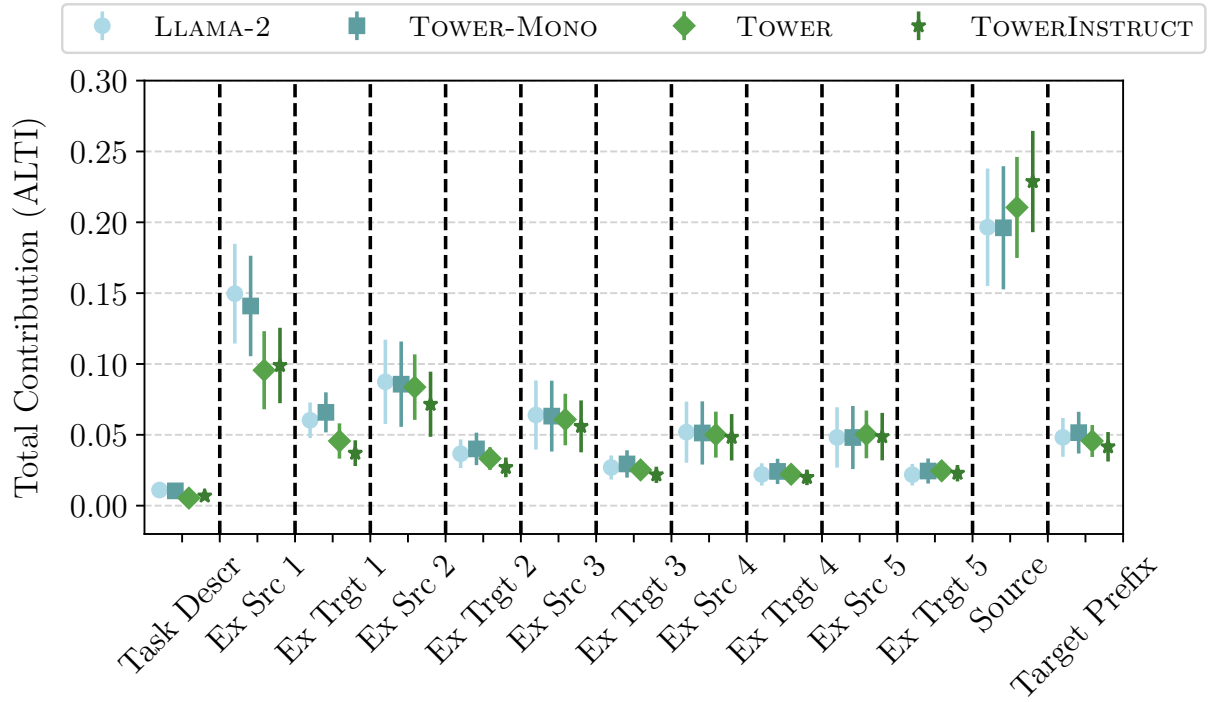


Figure 15: Illustration of context’s part-level contributions, when the task description is added. Translation direction: *Russian to English*

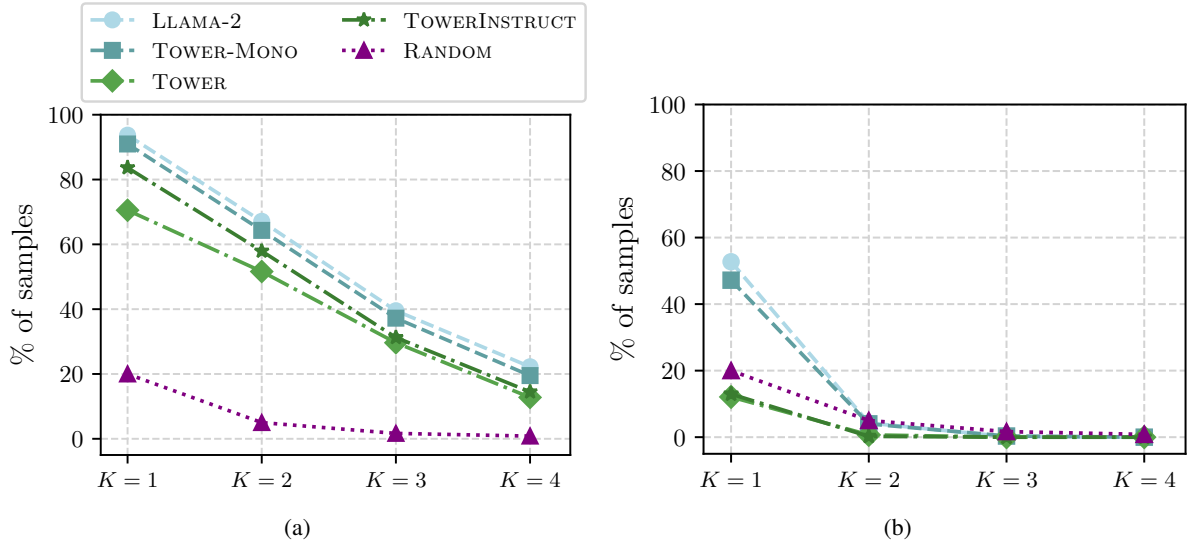


Figure 16: Proportion of en-de samples that follow positional bias, for different values of  $K$ , in the (a) original and (b) replace-last-ex settings.

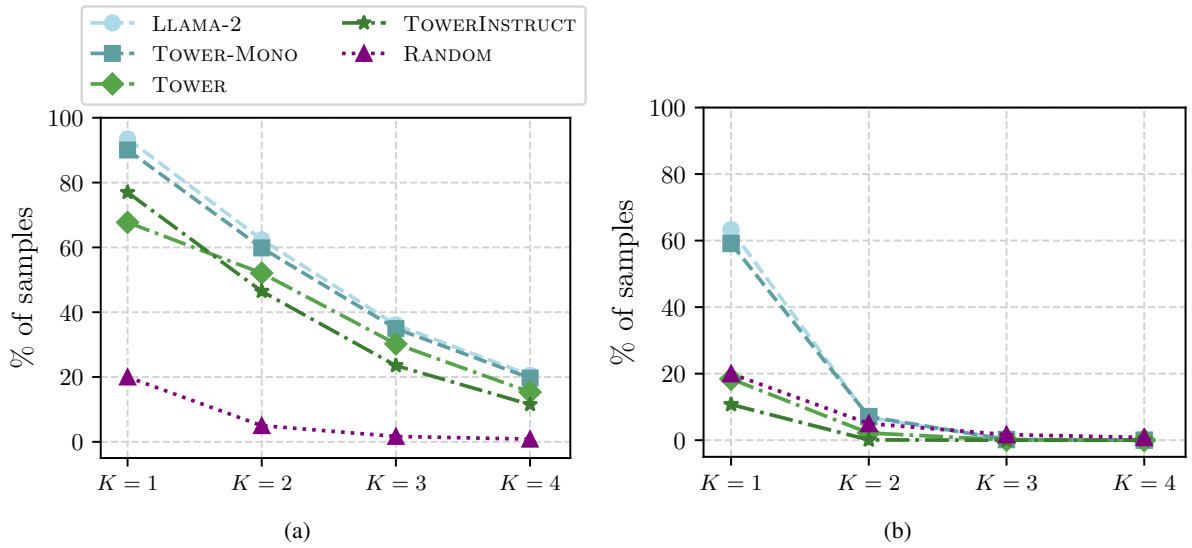


Figure 17: Proportion of ru-en samples that follow positional bias, for different values of  $K$ , in the (a) original and (b) replace-last-ex settings.

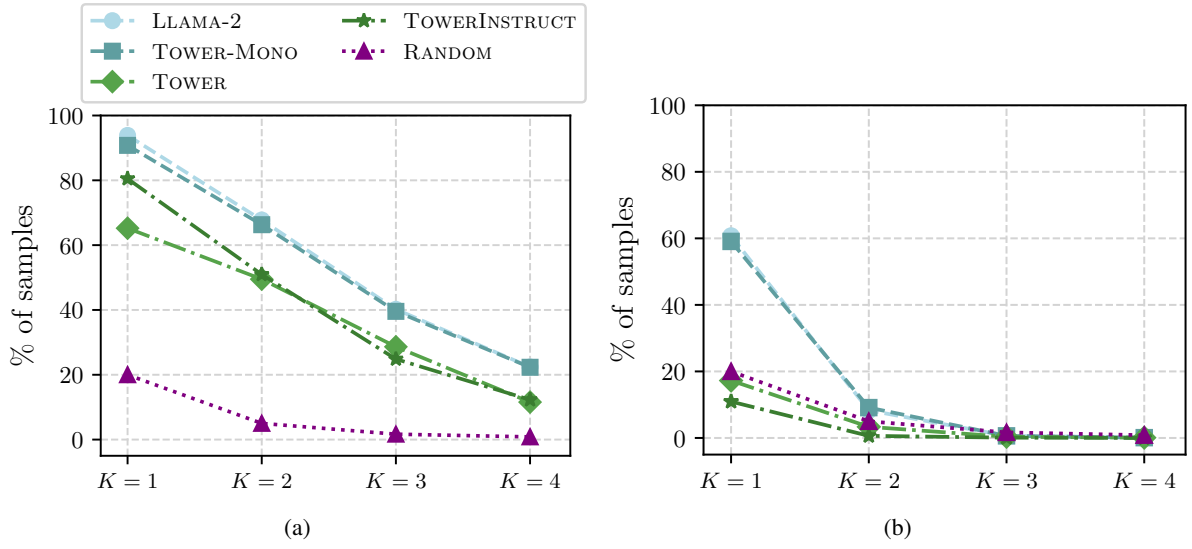


Figure 18: Proportion of en-ru samples that follow positional bias, for different values of  $K$ , in the (a) original and (b) replace-last-ex settings.

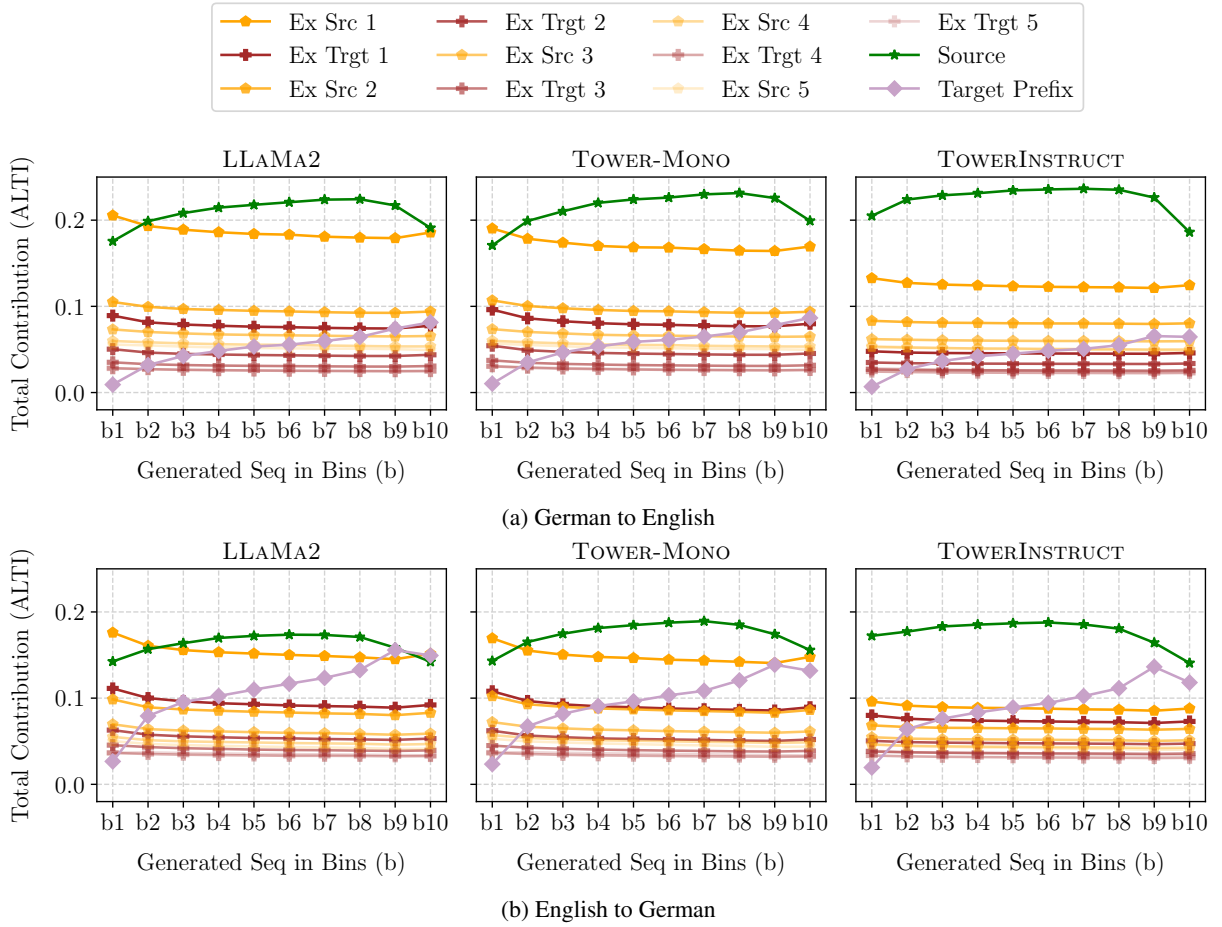


Figure 19: Illustration of how context contributions evolve across different generation stages, for the LLAMA-2, TOWER-MONO and TOWERINSTRUCT models. Each generated bin accounts for 10% of the generated sequence.

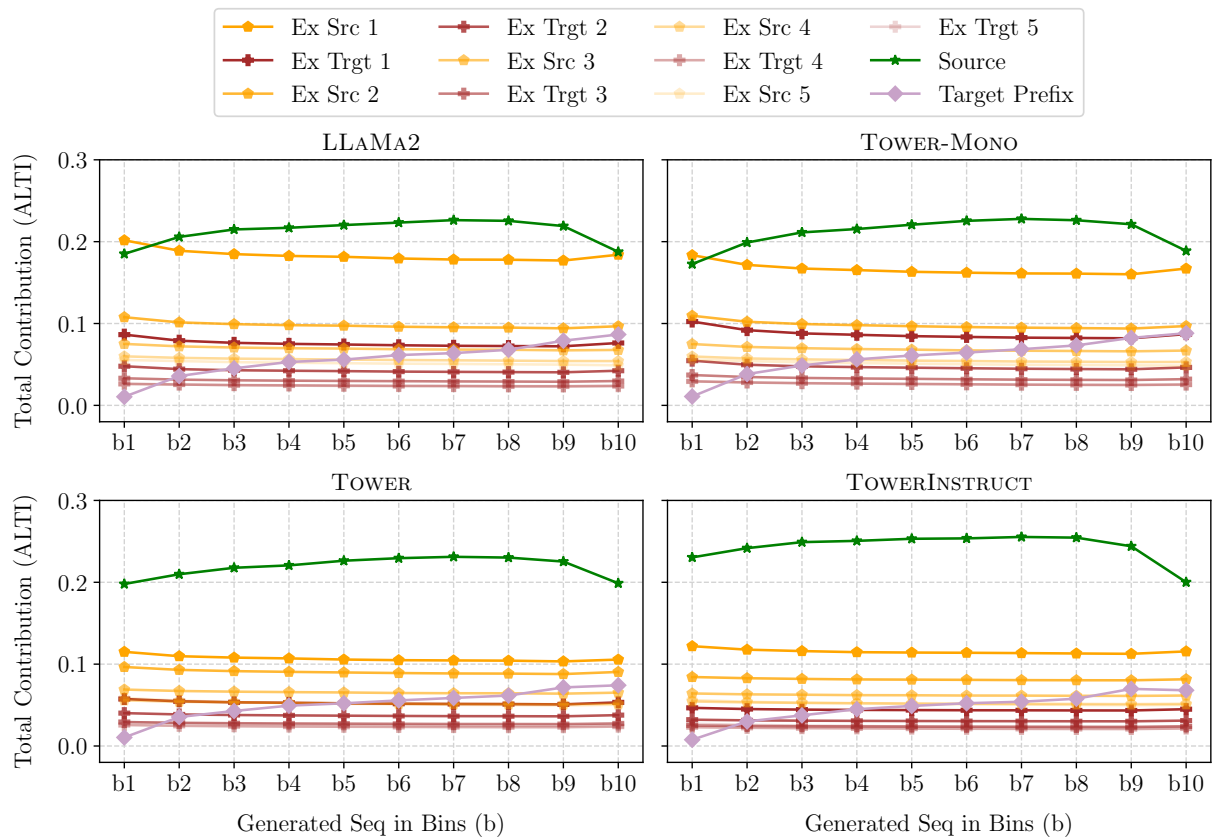


Figure 20: Illustration of how context contributions evolve across different generation stages, for all the examined models. Each generated bin accounts for 10% of the generated sequence. Translation direction: *Russian to English*

<b>E1   SRC</b>	Wie lange dauert es von Cefalù nach Taormina zu kommen?	
<b>E1   TGT</b>	How long does it take to get from Cefalù to Taormina?	
<b>E2   SRC</b>	Wie lange dauert es von Oslo nach Haugesund zu kommen?	
<b>E2   TGT</b>	How long does it take to get from Oslo to Haugesund?	
<b>E3   SRC</b>	Wie lange dauert es von Basel nach Montpellier zu kommen?	
<b>E3   TGT</b>	How long does it take to get from Basel to Montpellier?	
<b>E4   SRC</b>	Wie lange dauert es von Flensburg nach Århus zu kommen?	
<b>E4   TGT</b>	How long does it take to get from Flensburg to Århus?	
<b>E5   SRC</b>	Wie lange dauert es von Oslo nach Hammerfest zu kommen?	
<b>E5   TGT</b>	How long does it take to get from Oslo to Hammerfest?	
<b>SRC</b>	wie lange dauert es die gelben zu bestellen mit und ohne armlehne?	
<b>LLAMA-2</b> ✗		
<b>MT</b>	How long does it take to get from Oslo to Hammerfest?	
<b>TOWER</b> ✓		
<b>MT</b>	how long does it take to order the yellow with and without armrest?	

Table 8: Illustration of an example exhibiting anomalous source contribution for LLAMA-2 — which hallucinates, followed by TOWER’s contributions, which performs normally.



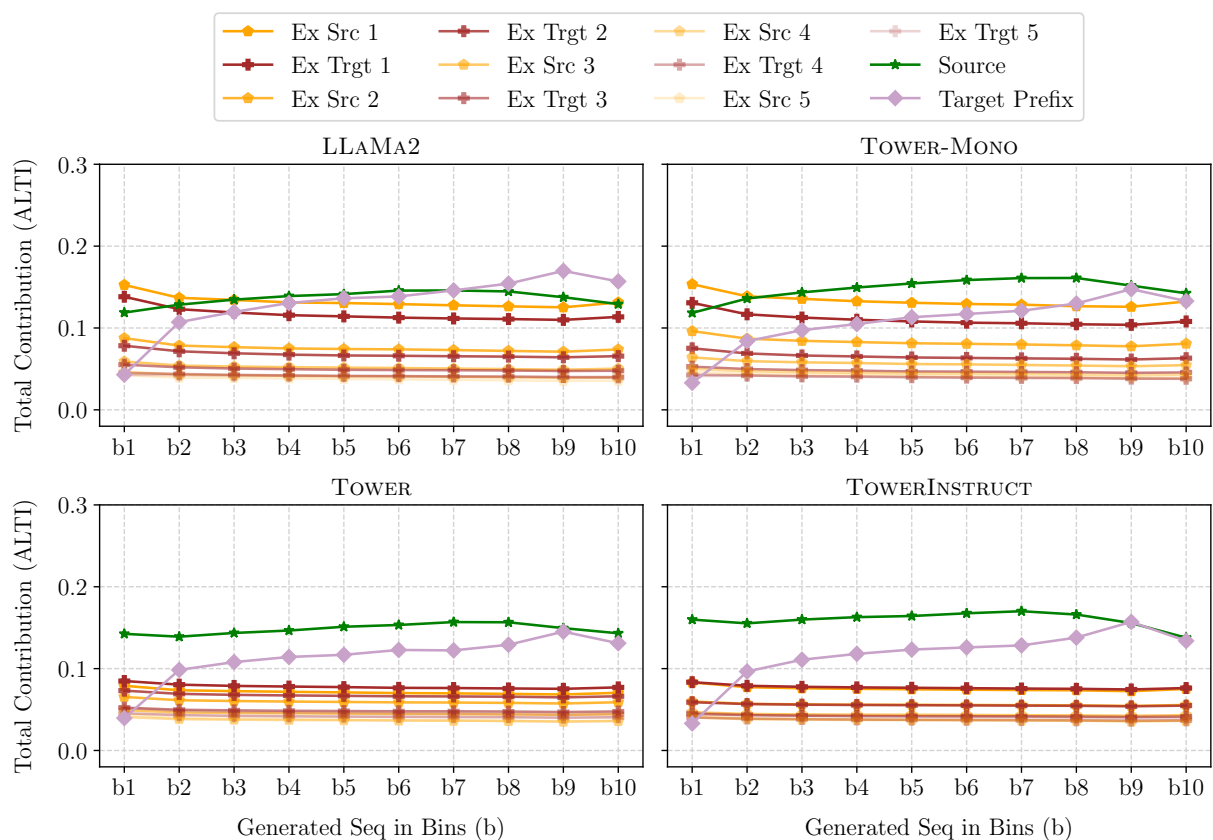


Figure 21: Illustration of how context contributions evolve across different generation stages, for all the examined models. Each generated bin accounts for 10% of the generated sequence. Translation direction: *English to Russian*

<b>E1 SRC</b>	Wir wünschen Ihnen einen angenehmen Aufenthalt in Maribor.
<b>E1 TGT</b>	We wish you a pleasant stay in Maribor.
<b>E2 SRC</b>	Wir wünschen Ihnen einen angenehmen Aufenthalt in Olomouc.
<b>E2 TGT</b>	We wish you a pleasant stay in Olomouc.
<b>E3 SRC</b>	Wir wünschen Ihnen einen angenehmen Aufenthalt in Debrecen.
<b>E3 TGT</b>	We wish you a pleasant stay in Debrecen.
<b>E4 SRC</b>	Wir wünschen Ihnen einen angenehmen Aufenthalt in Poznan.
<b>E4 TGT</b>	We wish you a pleasant stay in Poznan.
<b>E5 SRC</b>	Busbud hilft Ihnen, einen Bus von Lübeck nach Wismar zu finden.
<b>E5 TGT</b>	Busbud helps you find a bus from Lübeck to Wismar.
<b>SRC</b>	Wir verraten Ihnen, wo Sie im Raum Lübeck doch noch einen Weihnachtsbraten herbekommen.
<b>LLAMA-2 ✓</b>	
<b>MT</b>	We tell you where you can still get a Christmas roast in the Lübeck area.
<b>TOWER ✗</b>	
<b>MT</b>	Busbud helps you find a bus from Lübeck to Wismar.

Figure 22 displays two line charts illustrating the evolution of context contributions (ALTI) across different generation stages (1 to 10) for two models: LLAMA-2 and TOWER. The Y-axis represents the Total Contribution (ALTI), ranging from 0.0 to 0.4. The X-axis represents the Generated Seq in Bins (b).

The legend identifies the following series:

- Ex Src 1 (Orange line with circles)
- Ex Trgt 1 (Red line with circles)
- Ex Src 2 (Yellow line with circles)
- Ex Trgt 2 (Dark red line with circles)
- Ex Src 3 (Light orange line with circles)
- Ex Trgt 3 (Pink line with circles)
- Ex Src 4 (Light yellow line with circles)
- Ex Trgt 4 (Light pink line with circles)
- Ex Src 5 (Very light orange line with circles)
- Ex Trgt 5 (Very light pink line with circles)
- Source (Green line with circles)
- Target Prefix (Purple line with circles)

The charts show that the Source contribution (green line) is generally the highest across all models and bins, while the Target Prefix contribution (purple line) is the lowest. The contributions of the other sources and targets are relatively stable and low across the bins.

Table 9: Illustration of an example exhibiting anomalous source contribution for TOWER — which hallucinates, followed by LLAMA-2’s contributions, which performs normally.

<b>E1   SRC</b>	Telefónica Deutschland hat den SABRE Award EMEA gewonnen.
<b>E1   TGT</b>	Telefónica Deutschland has won the SABRE Award EMEA.
<b>E2   SRC</b>	New York City (Bundesstaat New York, USA): Promenade im Central Park.
<b>E2   TGT</b>	New York city (New York State, USA): Promenade in Central Park.
<b>E3   SRC</b>	New York City FC oder New England Revolution
<b>E3   TGT</b>	New York City FC or New England Revolution
<b>E4   SRC</b>	25.08 02:30 LA Galaxy - Los Angeles FC (Fußball,Major League Soccer)
<b>E4   TGT</b>	25.08 02:30 LA Galaxy - Los Angeles FC (Calcio,Major League Soccer)
<b>E5   SRC</b>	FC Schalke 04 hat 2 von den letzten 3 Spiele gegen VfL Wolfsburg gewonnen
<b>E5   TGT</b>	FC Schalke 04 has won 2 out of their last 3 matches against VfL Wolfsburg
<b>SRC</b>	New York City FC hat zum ersten Mal den Titel in der Major League Soccer gewonnen.
<b>LLAMA-2 ✓</b>	
<b>MT</b>	New York City FC has won the title in the Major League Soccer for the first time.
<b>TOWER ✓</b>	
<b>MT</b>	New York City FC has won the title in the Major League Soccer for the first time.

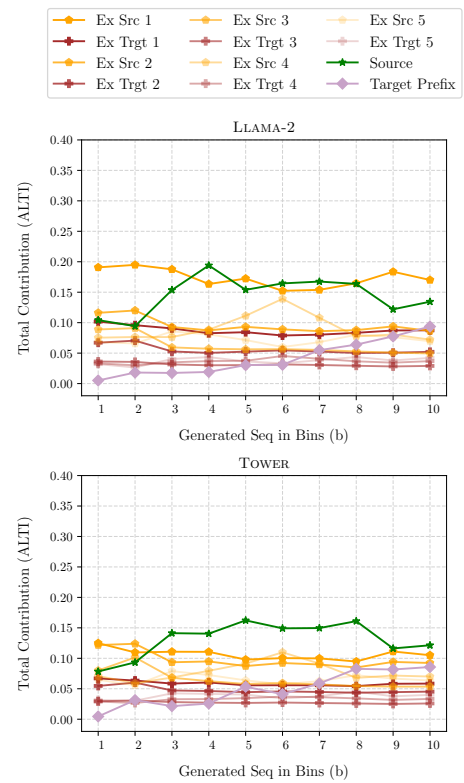


Table 10: Illustration of an example where both LLAMA-2 and TOWER produce correct translations. We observe that their contributions follow the average case trends for German to English translation.

<b>E1   SRC</b>	Arminia Bielefeld - Union Berlin2. Bundesliga.
<b>E1   TGT</b>	Arminia Bielefeld - Union Berlin2nd Bundesliga.
<b>E2   SRC</b>	Hertha BSC: Gewinner der 2. Bundesliga 2010/2011
<b>E2   TGT</b>	Hertha BSC: 2. Bundesliga winners 2010/2011
<b>E3   SRC</b>	Samstag, 9. März 2019 SV Darmstadt 98 Holstein Kiel
<b>E3   TGT</b>	Saturday, 9 March 2019 SV Darmstadt 98 Holstein Kiel
<b>E4   SRC</b>	Darmstadt Reisen von Saarbrücken nach Darmstadt in 4 stunden und 59 minuten
<b>E4   TGT</b>	Darmstadt Travel from Saarbrücken to Darmstadt in 4 hours and 59 minutes
<b>E5   SRC</b>	Das Wasser darf nicht heißer als 60 °C sein.
<b>E5   TGT</b>	The water must not be hotter than 60 °C.
<b>SRC</b>	Darmstadt 98 darf von der Rückkehr in die Fußball-Bundesliga träumen.
<b>LLAMA-2 ✓</b>	
<b>MT</b>	Darmstadt 98 can dream of returning to the Bundesliga.
<b>TOWER ✓</b>	
<b>MT</b>	Darmstadt 98 can dream of a return to the Bundesliga.

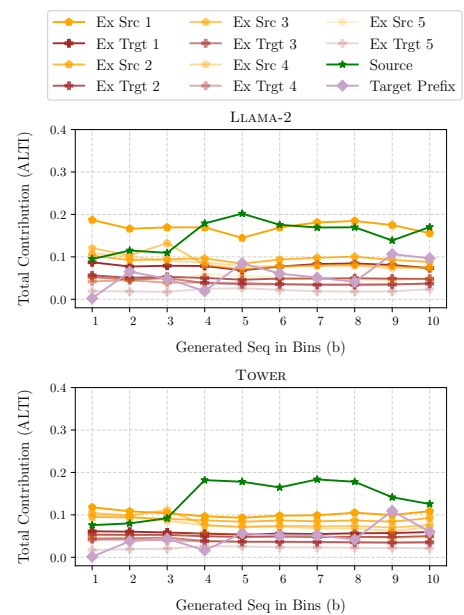


Table 11: Illustration of an example where both LLAMA-2 and TOWER produce correct translations. We observe that their contributions follow the average case trends for German to English translation.