

Scaling Attention Head Analysis via Gradient-Based Attribution in Context-Aware Machine Translation

Paweł Mąka and Yusuf Can Semerci and Jan Scholtes and Gerasimos Spanakis

Department of Advanced Computing Sciences

Maastricht University

{pawel.maka, y.semerci, j.scholtes, jerry.spanakis}
@maastrichtuniversity.nl

Abstract

In this paper, we introduce a gradient-based head attribution strategy where the Token-level Max-Margin loss is backpropagated to the attention maps. This framework enables a large-scale causal analysis of attention heads, making it suitable for LLMs. We evaluate our method on the task of disambiguation in Context-aware Machine Translation, where we analyze 50 phenomena across 4 models and 4 language directions. We empirically show the alignment of our method with the effects of increasing the attention scores of token-to-token relations on three models and two language directions, ensuring the robustness of our method. Our analysis reveals the presence of the "general-purpose" attention heads that improve the model's performance when attending to different relations. We find that the average attention a head assigns to a relation does not necessarily relate to the model's performance, which suggests that the models developed redundancies during training in terms of the head functions.

1 Introduction

In Context-Aware Machine Translation (MT), providing the system with surrounding sentences allows it to maintain document-level coherence and resolve linguistic ambiguities (Agrawal et al., 2018; Bawden et al., 2018; Voita et al., 2019b). While previous works often focused on specialized architectures (Tu et al., 2017; Bawden et al., 2018; Miculicich et al., 2018; Maruf et al., 2019; Huo et al., 2020; Zheng et al., 2021) or the encoder-decoder Transformer models, often referred to as the single-encoder architecture (Vaswani et al., 2017; Sun et al., 2022; Majumde et al., 2022; Gete et al., 2023; Post and Junczys-Dowmunt, 2023; Mohammed and Niculae, 2024),

the research focus has recently shifted toward the use of Large Language Models (LLMs) (Wang et al., 2023b, 2024; Alves et al., 2024; Cui et al., 2024; Wu et al., 2024), which utilize the decoder-only Transformer architecture to process context autoregressively. Despite the performance of these models (Pang et al., 2025), their size poses a challenge to analyzing their internal behavior.

Research on interpretability has demonstrated that specific attention heads within Transformers often learn to perform distinct, human-interpretable functions (Vig and Belinkov, 2019; Clark et al., 2019; Voita et al., 2019c; Olsson et al., 2022; Li et al., 2023). Techniques such as Activation Patching have been widely used to identify which components are causally responsible for specific model behaviors (Wang et al., 2023a; Zhang et al., 2025). Related to this, Modifying Heads (Mąka et al., 2025a), which is a technique where attention scores for specific linguistic relations are artificially shifted, has emerged as a tool to measure the causal importance of head-relation pairs. By forcing a head to attend to a specific contextual cue (e.g., a pronoun-antecedent relation), we can observe the impact on model performance as indicated by the accuracy in contrastive evaluations (Müller et al., 2018; Lopes et al., 2020).

Nevertheless, the computational cost of these causal interventions is significant because in modern LLMs with dozens of layers, there exist thousands of potential head-relation pairs. Consequently, performing a full sweep of modifications to identify functionally important components is prohibitively expensive. For instance, evaluating the EuroLLM 1.7B model (Martins et al., 2025) across all relevant relations requires thousands of individual evaluation runs. This limits our ability to scale the intervention-based studies to larger models or to include more language directions and linguistic phenomena.

In this paper, we propose an efficient framework

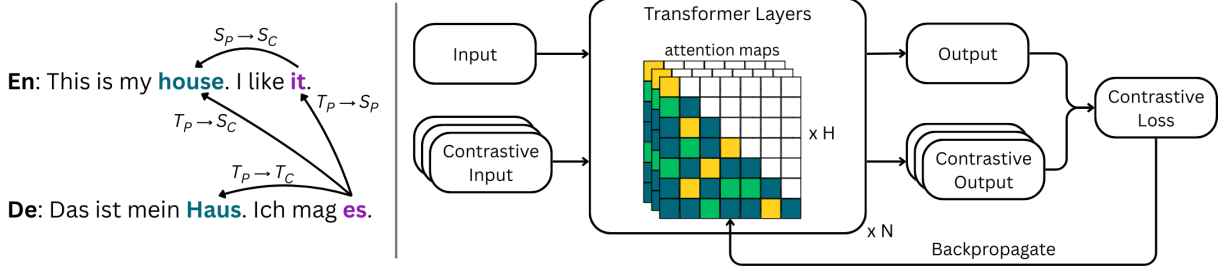


Figure 1: The token-to-token relations that we consider in this work (left). The illustration of the proposed strategy (right).

that automates the head-relation attribution, which extends the scalability of the interpretability studies for LLMs. Our method approximates the effects of head modification using gradient attribution. By backpropagating a loss into the attention maps (not to model’s parameters or inputs to the network), we can identify promising heads without the need to exhaustively evaluate each head-relation pair (see Figure 1 for an overview). We introduce and evaluate the Token-level Max-Margin loss $\mathcal{L}_{TMM}$, designed to align the gradient signal with the change in the accuracy of the contrastive evaluation after attention intervention. This allows us to select only the most promising (top- k) head-relation pairs for modification, significantly reducing the computational cost while maintaining the robustness of our analysis. We conduct a large-scale evaluation by analyzing 50 contextual phenomena across four language directions in four open-weights instruction-tuned LLMs: EuroLLM 1.7B (Martins et al., 2025) and 9B (Ramos et al., 2026), Qwen 2.5 1.5B (Yang et al., 2024; Qwen et al., 2025), and Gemma 3 1B (Team et al., 2025). We study the following contextual phenomena: Gender, Formality, Auxiliary, and Inflection (see Appendix A for more details). While we concentrate our evaluation on the task of Context-aware MT, the technique can, in principle, be applied to any task where a plausible token-to-token relation can be defined, which we test on the task of Indirect Object Identification (Wang et al., 2023a).

Our contributions are as follows:

- We propose a gradient-attribution framework to automatically identify salient head-relation pairs responsible for context utilization.
- We demonstrate that focusing on top- k heads identified via gradients preserves the accuracy of the analysis while reducing computational requirements by more than 96%.
- We perform a large-scale analysis of four

LLMs across four language directions. We confirm the presence of attention heads in LLMs that not only attend to linguistically relevant token-to-token relations but also improve the performance when the attention is increased.

- We discover attention heads that exhibit multifunctional behavior (attending to multiple specific relations), and we link this overlap with the generalizability of the models, which could open possibilities for improved transfer.

- We demonstrate a dissociation between the average attention score and the model’s performance, suggesting that LLMs develop functional redundancies.

2 Related Work

Historically, many dedicated architectures have been proposed for Context-aware MT (Miculicich et al., 2018; Voita et al., 2019b,a; Bao et al., 2021; Chen et al., 2022; Feng et al., 2022; Bulatov et al., 2022) including the popular multi-encoder in which a separate encoder is responsible for processing the context sentences (Jean et al., 2017; Miculicich et al., 2018; Maruf et al., 2019; Huo et al., 2020; Zheng et al., 2021), but the standard Transformer model with the sentences being concatenated, known as single-encoder (Vaswani et al., 2017; Tiedemann and Scherrer, 2017; Ma et al., 2020; Zhang et al., 2020), exhibited high performance despite its relative simplicity (Majumde et al., 2022; Sun et al., 2022; Gete et al., 2023; Post and Junczys-Dowmunt, 2023). Because of model scaling and pretraining on massive-scale datasets, the decoder-only LLMs have achieved state-of-the-art results in Context-aware MT (Wang et al., 2023b, 2024; Alves et al., 2024; Cui et al., 2024; Kocmi et al., 2024; Wu et al., 2024; Pang et al., 2025). However, the scale of the models presents a challenge for analyzing how the internal mechanisms in the

models relate to their performance on the end task.

Research has shown that the general-purpose translation metrics, such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020), do not measure how well the models use context (Hardmeier, 2012; Wong and Kit, 2012). This is because often only a few tokens require context in a given sentence (e.g., pronouns in a gendered language), and the standard metrics are insensitive to such small mistakes. To mitigate this issue, both contrastive (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019b; Lopes et al., 2020) and generative (Wicks and Post, 2023) evaluation datasets have been introduced.

Numerous methods aimed at explaining Natural Language Processing models have been proposed (Bau et al., 2018; Toneva and Wehbe, 2019; Ferrando et al., 2022; Langedijk et al., 2024; Meng et al., 2024). While using raw attention scores as explanations is a topic of debate (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019), attention remains the primary mechanism in the Transformer model enabling contextualization. Therefore, understanding its influence on the model’s behavior is an active area of research (Wiegrefe and Pinter, 2019; Abnar and Zuidema, 2020; Kobayashi et al., 2020, 2021; Bogoychev, 2021; Gheini et al., 2021; Mohebbi et al., 2023). Several works suggest that some heads exhibit specific functions (Clark et al., 2019; Voita et al., 2019c; Olsson et al., 2022; Todd et al., 2024) often linked to linguistically plausible relations (Vig and Belinkov, 2019; Tenney et al., 2019; Jo and Myaeng, 2020).

Similar to our approach, many works on interpretability employed gradients to explain the decisions of the models (Bach et al., 2015; Shrikumar et al., 2017; Sundararajan et al., 2017; Yin and Neubig, 2022) with attention heads being the focus of some of them (Selvaraju et al., 2017; Chefer et al., 2021; Song et al., 2024; Jo et al., 2025). The ultimate goal of those works was to explain the models by identifying the parts of the inputs that influence the model’s behavior (local explanations). Conversely, our work concentrates on identifying parts of the model itself (i.e., attention heads) that are influencing the final prediction, known as global explanation (Danilevsky et al., 2020; Madsen et al., 2022).

Attention Patching (Wang et al., 2023a; Zhang et al., 2025) and Modifying Heads (Mağa et al., 2025a) are the intervention-based interpretability

techniques that aim to identify meaningful functions of attention heads. While Attention Patching replaces the attention map for a head while processing the contrastive input with the saved attention map of the correct input, the Modifying Heads technique changes the attention to match the specified value C for a specific token-to-token relation, maintaining the distribution of attention for other tokens. It necessitates the specification of relations of interest, but allows for measuring the potentially positive effect of increased attention and enables the targeted fine-tuning of attention heads (Mağa et al., 2025a).

3 Background

In this section, we describe the existing concepts upon which our work builds. We start by formulating the attention mechanism in the Transformer model (Vaswani et al., 2017). Next, we present the attention intervention from the Modifying Heads method (Mağa et al., 2025a). Lastly, we describe the contrastive evaluation in Context-aware MT (Müller et al., 2018; Lopes et al., 2020).

3.1 Attention Mechanism

In the Transformer architecture, at each timestep, the multi-head self-attention is responsible for gathering the information from all input tokens. It consists of a specified number of heads that perform self-attention, and their outputs are concatenated and linearly projected to form the final output of the attention module. First, each head maps the input representations into the Q , K , and V projections by a learned linear transformation. Second, the scaled dot product between Q s and K s is calculated:

$$H^{l,h} = \frac{Q^{l,h} K^{l,hT}}{\sqrt{d_k}}, \quad (1)$$

where $\sqrt{d_k}$ is the scaling factor, and d_k is the dimension of the Q and K representations. Next, the attention scores $S^{l,h}$ are obtained by applying the softmax function to $H^{l,h}$:

$$S^{l,h} = \text{Softmax} \left(H^{l,h} \right). \quad (2)$$

Lastly, the output of head h in l -th layer is calculated as the sum of V representations weighted by the attention scores.

3.2 Modifying Heads

Modifying Heads (Mąka et al., 2025a) adjusts the attention scores corresponding to a specified token-to-token relation $\mathcal{X} \rightarrow \mathcal{Y}$. It makes the total attention score of each token in $\mathcal{X}$ given to $\mathcal{Y}$ tokens equal a desired value C while not changing the pre-softmax attention scores of all other tokens. It is formulated as follows:

$$\tilde{H}_{i,j}^{l,h} = \log \left(\frac{C}{|\mathcal{Y}|(1-C)} \sum_{k \in Y \setminus \mathcal{Y}} \exp \left(H_{i,k}^{l,h} \right) \right) \quad (3)$$

$\forall i \in \mathcal{X}, j \in \mathcal{Y},$

where $\tilde{H}$ represents the updated pre-softmax attention scores, H are the original pre-softmax scores, $k \in Y \setminus \mathcal{Y}$ are the attended tokens not present in the subset of interest $\mathcal{Y}$. Mąka et al. (2025a) modified the heads to different values of C for relations from ambiguous words to context cues and observed the changes in the accuracy of the models in contrastive evaluation. Intuitively, the method reveals how the model reacts to the changes in the function of one of its components. There are three notions of a head’s function: (1) the head is attending a relation, which can be measured by the average attention score; (2) the head attending a relation is necessary for the end task, measured by the drop of performance when the head is modified to $C = 0.01$; (3) the head attending a relation is beneficial to the task, indicated by the increase in the performance on the end task when modified to $C = 0.99$. In our work, we focus on the effects of Modifying Heads to $C = 0.99$, as it can lead to real-world improvements in the models through the targeted fine-tuning Mąka et al. (2025a).

3.3 Contrastive Evaluation

Contrastive evaluation assesses whether the model effectively uses the contextual information and is based on the contrastive datasets. It presents the model with the source sentence, source and target contexts, and several versions of the target sentence with only a single word being different in each of them. The examples are chosen to make the correct version dependent on the context. The model assigns the probability to each version, which is calculated from the model’s outputs with the whole sequence (including the target sentence) as input. If the right version is assigned the highest probability, the model is said to be correct. The

final score is the accuracy A of the model over the whole dataset.

In our study, we use two contrastive datasets: ContraPro (Müller et al., 2018) and Large Contrastive Pronoun Testset (LCPT; Lopes et al., 2020), which target pronoun disambiguation in English-to-German and English-to-French, respectively. To expand the analysis to different language phenomena, we adapt the ctxPro dataset (Wicks and Post, 2023) to contrastive evaluation. The original version is used for generative evaluation, meaning that the model generates the translation, and the script checks whether it contains the expected phrase. We generate the contrastive examples by replacing the expected word in the gold translation with a word sampled from all expected words for a particular linguistic phenomenon. Because this simple method does not ensure that the replacement word is morphologically plausible, we set the number of contrastive examples to maintain the difficulty of the test. We provide more details about the process in Appendix A.

4 Approach

Our goal is to approximate the effects of Modifying Heads to $C = 0.99$ for a specified token-to-token relation $\mathcal{X} \rightarrow \mathcal{Y}$ on the accuracy in the contrastive evaluation by using gradient attribution. For each example d in the dataset $\mathcal{D}$, we perform the forward pass of both the correct X^c and all incorrect X^{w_i} inputs. Next, we calculate the loss $\mathcal{L}$ based on the outputs of the model and backpropagate it into the attention scores $S^{l,h}$ of each head in each layer¹. We extract the gradients of the attention scores corresponding to the specified relation and aggregate them over the dataset:

$$G_{\mathcal{X} \rightarrow \mathcal{Y}}^{l,h} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \max_{i \in \mathcal{X}^d, j \in \mathcal{Y}^d} \left(- \left(\nabla_{S_c^{l,h}} \mathcal{L}^d \right)_{ij} \right), \quad (4)$$

where $\nabla_{S_c^{l,h}}$ represents the gradient with respect to the attention scores $S_c^{l,h}$ of the correct contrastive example, and $\mathcal{L}^d$ is the loss of the d -th example from the contrastive dataset $\mathcal{D}$. For the multi-token phrases and contextual cues, we aggregate the maximum gradient over the token pairs, as this functionally corresponds to a head attending any part of the cue from any part of the phrase.

¹It should be noted that we do not use the gradients of the model’s parameters but of the attention maps directly.

The choice of the loss function $\mathcal{L}$ determines the quality of the resulting gradients, as we explore in Section 5.1. Motivated by the contrastive evaluation, we propose to use max-margin loss based on contrastive tokens, which we call Token-level Max-Margin loss (TMM):

$$\mathcal{L}_{TMM}^d = \max \left(0, \mu - \bar{P}_\theta(x^{d,c}) + \max_i \left(\bar{P}_\theta(x^{d,w_i}) \right) \right), \quad (5)$$

where $\bar{P}_\theta(x)$ is the probability of the token x assigned by the model, and margin μ is the hyper-parameter. We select the lowest probability in case of multi-token contrastive words or phrases. This loss guides the model’s attention scores to increase the relative probability difference between contrastive tokens only in cases where the model is wrong (assigns lower probability to the correct token) or the difference is smaller than μ .

While the backpropagation of the loss into the attention scores of each head is a good indicator of the difference in accuracy ΔA on the contrastive dataset when modified to $C = 0.99$, gradients are non-linear or even non-monotonic (see Section 5.2). For this reason, we still apply Modifying Heads to $C = 0.99$ and evaluate on the contrastive dataset, but only for head-relation pairs that exhibit the highest gradients (top- k). This greatly reduces the computational requirements while preserving the robustness of the measurement.

5 Experiments

In this section, we investigate the properties of our proposed method of gradient attribution of head-relation pairs. We adapt the notation of the relations from Mąka et al. (2025a) but extend the number of relations to include the relation tokens as the input to the network. This is due to the autoregressive nature of LLMs and the fact that the token identity is retained by the Transformer through the layers (Brunner et al., 2020). The relations investigated in this study are as follows (see also Figure 1 left):

- $S_P \rightarrow S_C$,
- $S_P \rightarrow S_{C+1}$,
- $S_{P+1} \rightarrow S_C$,
- $S_{P+1} \rightarrow S_{C+1}$,
- $T_P \rightarrow S_C$,
- $T_P \rightarrow S_{C+1}$,
- $T_P \rightarrow S_P$,
- $T_P \rightarrow S_{P+1}$,

- $T_P \rightarrow T_C$,
- $T_P \rightarrow T_{C+1}$,

where S_P, T_P represent the ambiguous phrases as output in the source and target language respectively, S_C, T_C represent the context cue as output (in the source and target language), and the tokens as the input are represented by a +1 (e.g., S_{C+1} is the context cue in the source language as the input to the network). Note that we do not include the relations from the ambiguous phrase in the target language as input, as at this timestep the model has already decided on the correct (or incorrect) phrase. We use the notation {layer}-{head} to identify attention heads (e.g., 3-16 represents the 16th head in the third layer of the model).

We perform a full sweep of Modifying Heads for all relevant relations for the EuroLLM 1.7B instruction-tuned model² (Martins et al., 2025) on the ContraPro contrastive dataset (Müller et al., 2018). In addition to the current source and target sentences, we concatenate 5 previous sentences on both the source and target side (context size of 5). Additionally, we test the losses on this model with the context size of one. Furthermore, we evaluate the methods on the Qwen 2.5 1.5B instruction-tuned model³ (Yang et al., 2024; Qwen et al., 2025) on ContraPro in English-to-German, LCPT in English-to-French (Lopes et al., 2020), and Auxiliary phenomenon in English-to-German from the contrastive version of the ctxPro dataset (Wicks and Post, 2023). Next, we evaluate the third model Gemma 3 1B instruction-tuned (Team et al., 2025)⁴ on ContraPro. The last model we test is the larger EuroLLM 9B Instruct⁵ (Ramos et al., 2026) on ContraPro. But, because of the prohibitive computational cost of evaluating each head-relation pair (see Appendix C), we examine the results of the smaller EuroLLM 1.7B model and select 5 relations that achieve the highest increase in accuracy ΔA ($T_P \rightarrow T_{C+1}$, $T_P \rightarrow T_{P+1}$, $T_P \rightarrow T_C$, $T_P \rightarrow S_{C+1}$, $T_P \rightarrow T_C$). The total number of evaluations are presented in Table 4, and the time and memory requirements are shown in Tables 5 and 6 for evaluation and calcu-

²<https://huggingface.co/utter-project/EuroLLM-1.7B-Instruct>

³<https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>

⁴<https://huggingface.co/google/gemma-3-1b-it>

⁵<https://huggingface.co/utter-project/EuroLLM-9B-Instruct-2512>

Loss	EuroLLM		EuroLLM 9B*	ContraPro	Qwen		Gemma
	Context 5	Context 1			LCPT	Auxiliary	
$\mathcal{L}_{NLL}$	0.0050	0.0036	0.0008	0.0201	0.0120	0.3562	0.0363
$\mathcal{L}_{logC}$	0.3851	0.4354	0.4624	0.4714	0.6219	0.4832	0.2568
$\mathcal{L}_{logTC}$	0.4438	0.5907	0.4665	0.5440	0.6254	0.5167	0.3140
$\mathcal{L}_{TC}$	0.4225	0.5092	0.3704	0.5127	0.6163	0.4970	0.3761
$\mathcal{L}_{logMM}$	0.4592	0.4883	0.5318	0.5542	0.6180	0.4500	0.3370
$\mathcal{L}_{logTMM}$	0.4846	0.5619	0.4809	0.5692	0.6578	0.4802	0.3448
$\mathcal{L}_{TMM}$	0.5362	0.5709	0.5606	0.5838	0.6799	0.4834	0.4513

Table 1: **Average precision** of finding the head-relation pairs that improve accuracy when modified to $C = 0.99$ compared to the unmodified model ΔA by more than the 99% confidence interval (of the unmodified model) for all relations of interest for all tested models and phenomena. *EuroLLM 9B was evaluated on the limited set of token-to-token relations.

lating gradients, respectively.

We compare the Token-level Max-Margin loss $\mathcal{L}_{TMM}$ (eq. 5) with the following losses as baselines:

- Negative Log Likelihood of the ambiguous token:

$$\mathcal{L}_{NLL}^d = -\log\left(\bar{P}_\theta(x^{d,c})\right), \quad (6)$$

- Contrastive loss for the whole sequence:

$$\mathcal{L}_C^d = -\bar{P}_\theta(X^{d,c}) + \max_i \left(\bar{P}_\theta(X^{d,w_i}) \right), \quad (7)$$

- Contrastive loss for the ambiguous token - similar to Yin and Neubig (2022),

$$\mathcal{L}_{TC}^d = -\bar{P}_\theta(x^{d,c}) + \max_i \left(\bar{P}_\theta(x^{d,w_i}) \right), \quad (8)$$

- Max-Margin loss for the whole sequence:

$$\mathcal{L}_{MM}^d = \max\left(0, \mu - P_\theta(X^{d,c}) + \max_i \left(P_\theta(X^{d,w_i}) \right)\right). \quad (9)$$

Additionally, we include log-space versions of contrastive losses, where probabilities assigned by the model are replaced with their logarithms. We note the losses as $\mathcal{L}_{logX}$ (e.g., $\mathcal{L}_{logTC}$ for the logarithm version of the contrastive loss for the ambiguous token $\mathcal{L}_{TC}$). Because of the issues with numerical stability, we consider only log-probability versions of losses based on the whole sequence ($\mathcal{L}_{logC}$ and $\mathcal{L}_{logMM}$). We set $\mu = 0$, which corresponds to calculating the loss only for the examples where the model was not correct.

For all tested losses, we aggregate gradients for all head-relation pairs and all examples in the contrastive dataset for which a particular relation is present (context cue inside the context size). We

measure Pearson r and Spearman ρ correlation coefficients between the gradients and the accuracy difference ΔA between the unmodified model and the model with Modified Heads to $C = 0.99$ (for the same head-relation pair). Furthermore, we calculate the average precision scores, Normalized Discounted Cumulative Gain (NDCG), and ranks of retrieving head-relation pairs that result in the accuracy difference ΔA above the 99% confidence interval of the unmodified model. We present the hyperparameters of evaluation in Appendix B.

5.1 Gradient-Modifying Alignment

We present the baseline results for all models and phenomena in Table 7 in Appendix F. The results in terms of Pearson r and Spearman ρ correlation coefficients for the tested models can be seen in Table 8 in Appendix F. We additionally include the correlations of two subsets of head-relation pairs: those with an accuracy difference greater than zero ($\Delta A > 0$) and the remaining ones ($\Delta A \leq 0$). This is motivated by the fact that fundamentally different mechanisms are responsible for both effects. While the model’s positive response to Modifying Heads to $C = 0.99$ can be interpreted precisely (the attention score for a particular head and relation is increased), the negative response can be attributed to a reduction in the attention score for another token that is important for the head. This can be observed in Figure 9 in Appendix F, where we show the comparison between the aggregated gradients and the accuracy difference ΔA for the EuroLLM 1.7B Instruct model. Token-level Max-Margin loss $\mathcal{L}_{TMM}$ shows the highest Spearman correlations ρ for accuracy differences greater than zero ($\Delta A > 0$) for most of the tested scenarios.

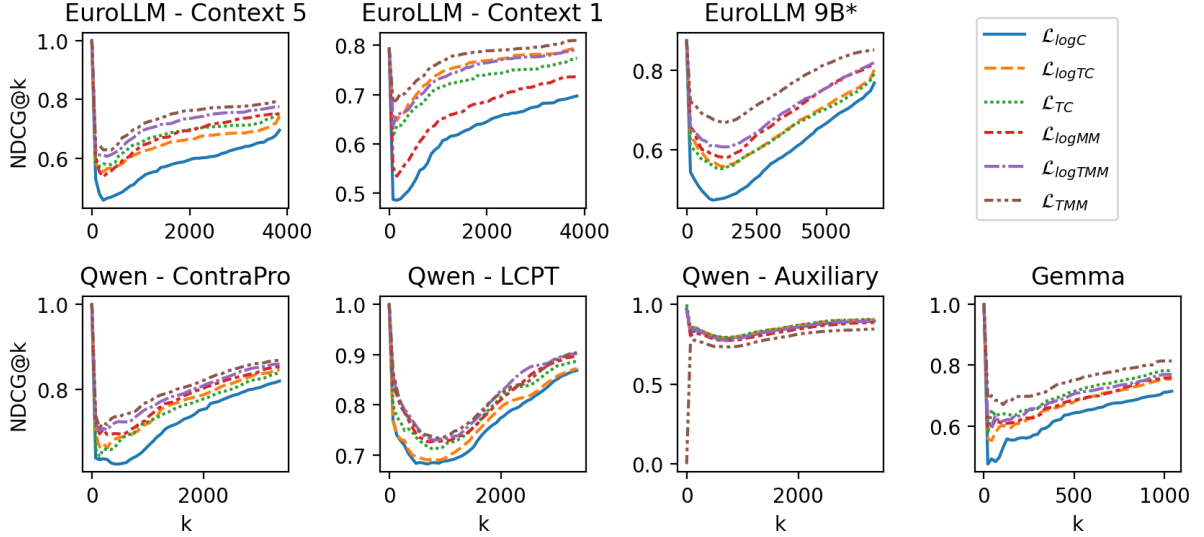


Figure 2: The **Normalized Discounted Cumulative Gain (NDCG)** of finding the head-relation pairs that improve accuracy when modified to $C = 0.99$ compared to the unmodified model ΔA by more than the 99% confidence interval (of the unmodified model) for tested losses. *EuroLLM 9B was evaluated on the limited set of token-to-token relations.

The exceptions were the Qwen model on Auxiliary and LCPT datasets.

Table 1 shows the results in terms of the average precision of retrieving the head-relation pairs that improve the accuracy ΔA by more than 99% confidence interval of the unmodified model. Token-level Max-Margin loss $\mathcal{L}_{TMM}$ achieves the best results in most of the cases, with the exceptions of EuroLLM 1.7B with context size of 1 (where it achieved the second best score), and Qwen on the Auxiliary task. In Figure 2 we show the Normalized Discounted Cumulative Gain (NDCG) for all tested methods, where the proposed loss achieved the best results for most scenarios. The exceptions are Qwen on LCPT and Auxiliary datasets. Finally, we examined the ranks of head-relation pairs in terms of the accuracy difference ΔA given by the backpropagated losses. The results can be seen in Figure 3. Ranking according to the Token-level Max-Margin loss $\mathcal{L}_{TMM}$ outperforms other losses for most of the tested models and phenomena.

While Token-level Max-Margin loss $\mathcal{L}_{TMM}$ outperforms other losses in most tested scenarios, it was surpassed on EuroLLM 1.7B with context size of 1 and Qwen on LCPT and Auxiliary datasets. We argue for using it for all models and phenomena, because it is still able to retrieve all relevant head-relation pairs in those scenarios within the top 1% of the ranking (see Figure 3).

5.2 Why Gradients Work?

While gradient-attribution methods have seen successes (Sundararajan et al., 2017; Yin and Neubig, 2022; Song et al., 2024; Jo et al., 2025), they have been criticized for their unreliability as explanations (Wang et al., 2020). In this section, we try to answer the question of why our method with Token-level Max-Margin loss $\mathcal{L}_{TMM}$ showed moderate to strong correlations with the accuracy difference ΔA and effectively ranked top head-relation pairs.

We calculate gradients for three selected heads when modified to values in $C \in \{0.01, 0.1, 0.2, \dots, 0.9, 0.99\}$. Doing this allows us to see the monotonicity of the gradient-accuracy increase relation. While the relation for the accumulated gradients (shown in Figure 4a) is monotonic (to the granularity of the tested values), the gradients for the individual examples (depicted in Figure 4b) are not only non-monotonic but also change sign. This finding highlights that gradient-attribution methods are hindered by the non-monotonicity of the underlying function, and the accumulation of the gradients can mitigate this issue (Sundararajan et al., 2017).

6 Analysis

In this section, we analyze the models on a wide range of language directions and phenom-

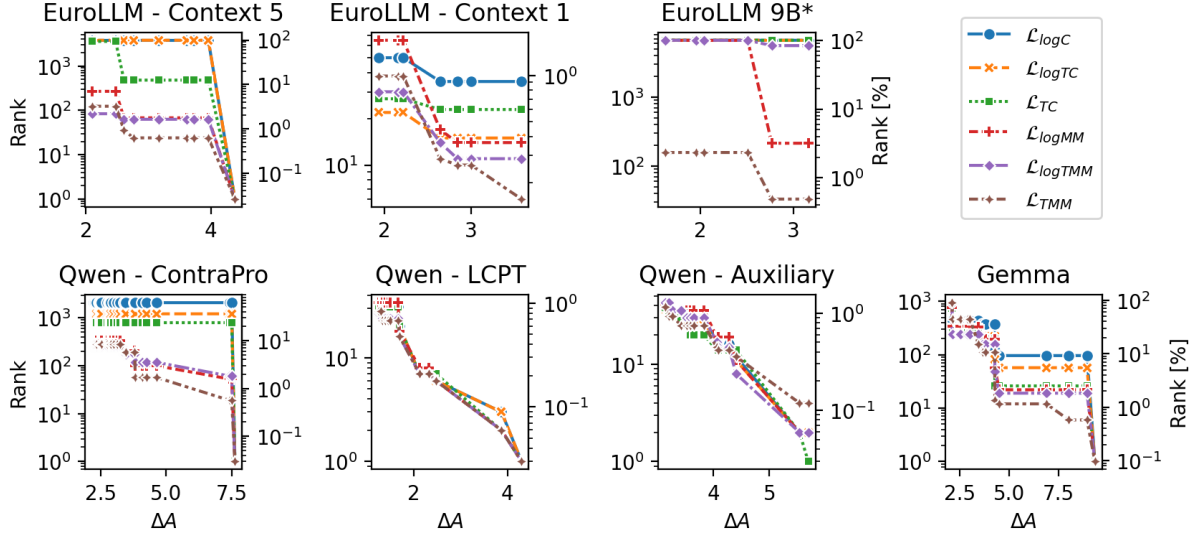


Figure 3: The **ranks** of head-relation pairs according to the gradients based on the tested losses for all tested models and phenomena. For clarity, we show the head-relation pairs that improve accuracy compared to the unmodified model ΔA by more than the 99% confidence interval (of the unmodified model). *EuroLLM 9B was evaluated on the limited set of token-to-token relations.

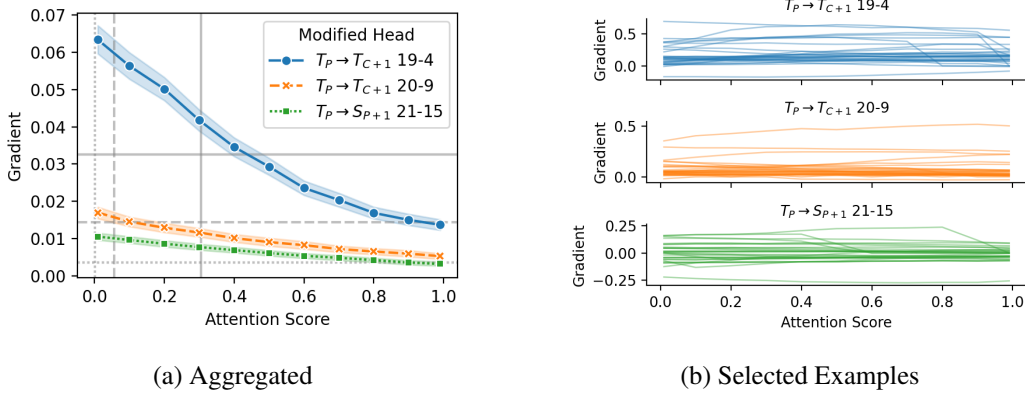


Figure 4: The gradients (based on the Token-level Max-Margin Loss $\mathcal{L}_{TMM}$) of selected heads when modified to different values of Attention Scores C when: aggregated (4a), and sampled 30 examples with non-zero gradient for most head-relation pairs and C values (4b).

ena. Specifically, we include translating from English to German, French, Polish, and Russian, and the following phenomena: Gender (we use ContraPro and LCPT for English-to-German and English-to-French, respectively, and ctxPro Gender for English-to-Polish and English-to-Russian), Auxiliary, Formality, and Inflection (the last one available only for English-to-Polish and English-to-Russian). This amounts to 50 scenarios across four models and four language directions.

We apply our gradient-based framework by first backpropagating the Token-level Max-Margin loss $\mathcal{L}_{TMM}$ and selecting top- k head-relation pairs in terms of aggregated gradient. Motivated by the

rank of top head-relation pairs (see Figure 3), we set the number of heads evaluated by Modifying Heads to be equal to 3% of all head-relation pairs. Note that, due to the lack of specified context cues in the ctxPro dataset for Formality and Inflection, we cannot define all of the relations explored previously for those phenomena, as only the contextually dependent phrases are identified. This shortens the list of relations to: $T_P \rightarrow S_P$ and $T_P \rightarrow S_{P+1}$, effectively reducing the number of head-relation pairs. For those phenomena, we still maintain the same computational budget for modifying top heads as for other phenomena. Therefore, we modify 15% of top head-relation

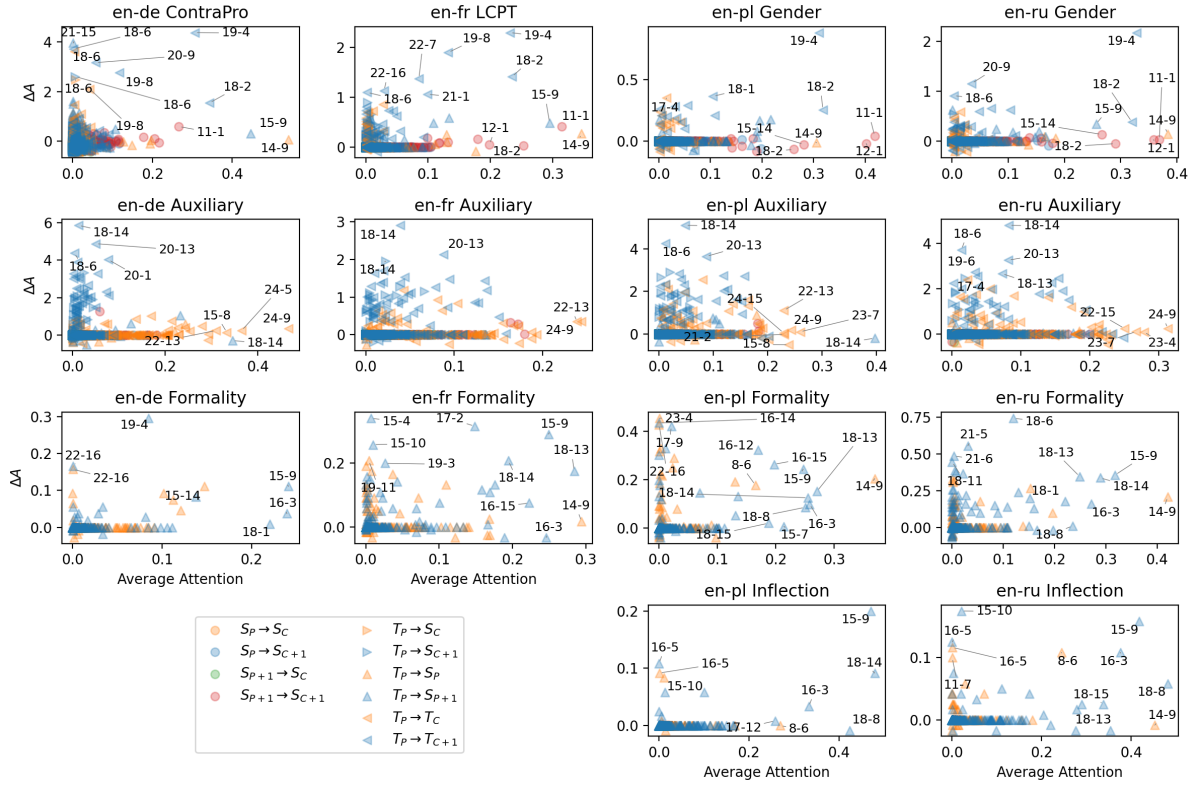


Figure 5: The results in terms of average attention scores (x axes) and accuracy increase ΔA (y axes) for head-relation pairs for **EuroLLM 1.7B Instruct** on all tested phenomena and language directions. To improve clarity, we omitted the head-relation pairs with a large negative ΔA .

pairs for Formality and Inflection. While gradients align well with the change in the accuracy ΔA , we saw that some heads can attend a relation but the model will not improve the performance when those heads are modified (e.g., 14-9 $T_P \rightarrow S_P$ in Figure 5). To analyze those heads, we modify 0.3% of head-relation pairs with the highest average attention scores. Because the evaluation of the EuroLLM 9B model on ContraPro was done for a limited set of relations, we additionally evaluate the top 3% head-relation pairs according to our method and analyze the combined results. We omit evaluating Formality and Inflection for this model due to the low statistical significance of the results for those phenomena on the smaller models.

In general, by focusing our computational resources exclusively on the promising head-relation pairs, we were able to scale the analysis to 50 phenomena across 4 different language directions and 4 models in total. This allowed us to assess the cross-lingual and cross-phenomena behavior of the models while reducing the computational cost of the analysis by more than 96%. The re-

sults can be found in Figures 5 (EuroLLM 1.7B Instruct), 10 (EuroLLM 9B Instruct), 11 (Qwen 2.5 1.5B Instruct), and 12 (Gemma 3 1B Instruct).

6.1 Results

Our work confirms that heads appear to have specific functions by attending to a specific relation or improving the model’s performance when modified to $C = 0.99$. While this is in line with previous research (Clark et al., 2019; Voita et al., 2019c; Olsson et al., 2022; Mąka et al., 2025a), our analysis reveals that the heads can attend to or benefit from modifying multiple relations. This generalization of capability can manifest in several forms spanning different types of relations:

- Same relation but the target token on the input and output - for example, the heads 19-8 in EuroLLM 1.7B and 24-11 in Qwen 2.5 1.5B improve accuracy on the ContraPro dataset when modified to attend both the $T_P \rightarrow T_C$ and $T_P \rightarrow T_{C+1}$ relations. This corroborates the finding that the identity of the input token is maintained through the layers (Brunner et al., 2020), but such heads are not widespread.

- Same relation and phenomenon but different language directions - the examples include head-relation pairs $14-9 T_P \rightarrow S_P$ in EuroLLM 1.7B, $32-3 T_P \rightarrow T_{C+1}$ in EuroLLM 9B, $16-2 T_P \rightarrow T_{C+1}$ in Gemma 3 1B for ContraPro, LCPT, and Gender (both for English-to-Polish and English-to-Russian), and $20-13 T_P \rightarrow T_{C+1}$ and $22-13 T_P \rightarrow T_C$ for Auxiliary (all language directions) in EuroLLM 1.7B, among others.

- Same relation but different language direction and phenomenon - for example, head-relation pairs $16-3 T_P \rightarrow S_{P+1}$ for Formality and Inflection (high attention but not accuracy difference) in EuroLLM 1.7B, and $12-2 T_P \rightarrow S_{P+1}$ in Gemma 3 1B exhibit this behavior.

- Different relation, language direction, and phenomenon - most notably, heads 18-6, and 18-14 in EuroLLM 1.7B, 25-4 in EuroLLM 9B, 20-2 in Qwen 2.5 1.5B, and 18-3 in Gemma 3 1B.

The existence of those "general-purpose" attention heads suggests that the updates to the residual stream coming from those heads share some component, as the information they encode can improve the model's performance on several tasks. We leave the detailed exploration of this finding for future work.

Importantly, the results for Formality and Inflection in terms of the accuracy difference ΔA when modified to $C = 0.99$ are not statistically significant (see Table 7 for the confidence intervals of the results of the unmodified models). We hypothesize that this can be caused by the limited set of relations we investigated (due to the lack of context cues being annotated by the ctxPro dataset). Additionally, modifying to $C = 0.99$ requires the model not to achieve a perfect score, as there needs to be room for improvement to observe any positive change in accuracy. This shows a limitation of the Modifying Heads technique: the need for well-defined relations between tokens and evaluation metrics that are challenging for the models.

6.2 Unresponsive Heads

Many heads show high average attention scores with no improvement in accuracy when modified to $C = 0.99$ for specific relations. In EuroLLM 1.7B, those include head-relation pairs $14-9 T_P \rightarrow S_P$ for Gender and Formality, $11-1 S_{P+1} \rightarrow S_{C+1}$ for Gender, $24-9 T_P \rightarrow T_C$ for Auxiliary, and $18-8 T_P \rightarrow S_{P+1}$ for Inflection (among others). We investigate whether this is due to those heads being

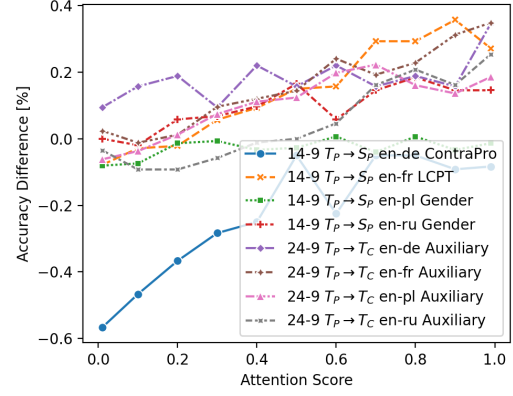


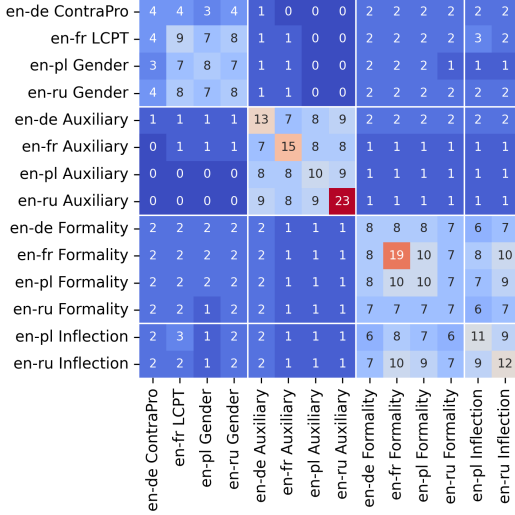
Figure 6: Accuracy difference ΔA of selected head-relation pairs for different values of modified attention scores C for EuroLLM 1.7B Instruct.

"optimal" in their apparent function of attending to the relations. To test this hypothesis, we modify selected heads (namely, $14-9 T_P \rightarrow S_P$ for Gender and $24-9 T_P \rightarrow T_C$ for Auxiliary in all language directions) to the values of C spanning the whole range: $C \in \{0.01, 0.1, 0.2, \dots, 0.9, 0.99\}$. If our hypothesis is true, we expect to see a plateau for C 's larger than the average attention score of the unmodified model.

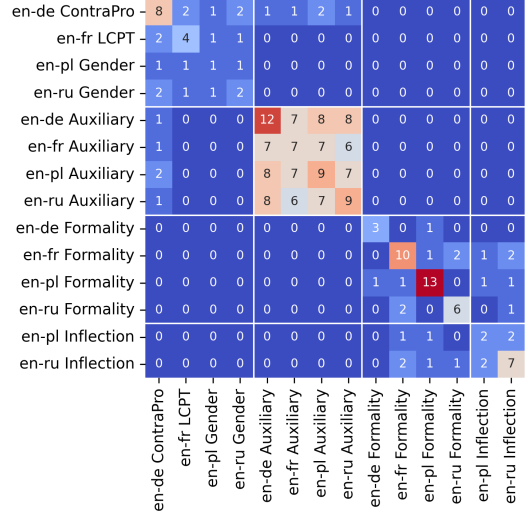
The results can be seen in Figure 6. While we do not see a statistically significant improvement above the average attention scores, we also do not observe the drop in accuracy ΔA below that point ($14-9 T_P \rightarrow S_P$ for ContraPro shows slightly larger drop in accuracy, but is still within the confidence interval of the unmodified model; see Table 7). This suggests that even though the heads in question attend to the relations, they are not crucial to the model for disambiguating the context-dependent phrases. This corroborates the notion that raw attention scores alone are not enough to interpret the models (Jain and Wallace, 2019). We hypothesize that the existing redundancies inside the model (other heads attending the same relation, such as $15-9 T_P \rightarrow S_{P+1}$ for Gender) enable the model to function properly.

6.3 Link to Generalizability

Motivated by the overlap of some heads between phenomena and language directions, we quantified the extent of shared head-relation pairs between each phenomena. We selected the head-relation pairs that exhibit an increase in accuracy ΔA greater than half of the maximum ΔA for each phenomenon and language direction sepa-



(a) Average Attention



(b) ΔA

Figure 7: Number of shared head-relation pairs between language directions and phenomena for EurolLM 1.7B Instruct. We select heads that exhibit average attention (7a) and ΔA (7b) higher than half of the maximum value for a specified language direction and phenomenon.

rately. We applied the same criterion to average attention scores. The results can be seen in Figure 7.

We observed a large overlap for Auxiliary between language directions for both the average attention scores and accuracy increases ΔA . For Gender (including ContraPro and LCPT), the overlap is high based on the average attention, and substantially lower based on the accuracy increase ΔA . Similar trends can be seen for Formality and Inflection, although the results in terms of accuracy increase ΔA are not statistically significant. We argue that the overlap based on the accuracy increase ΔA better explains the transfer during training between phenomena observed by Mąka et al. (2025b), where it was higher for Auxiliary, very limited inside other phenomena, and not present at all between different phenomena. Intuitively, positive changes to accuracy when modified indicate the potential of certain heads to improve the model’s performance during training. Whether these heads actually learn to attend more to the relations during training and how this relates to the improvements in other phenomena is left for future work.

6.4 Tuning Attention Heads

The Modifying Heads technique artificially intervenes in the model’s forward pass. In this section, we examine how this intervention, when

Model		BLEU	COMET	ContraPro	
Baseline		21.41	0.7862	83.14%	
Head	Relation	BLEU	COMET	Modified	Tuned
19-4	$T_P \rightarrow T_{C+1}$	21.33	0.7857	+2.84%	+2.72%
18-6	$T_P \rightarrow T_{C+1}$	21.19	0.7862	+3.00%	+2.25%
18-6	$T_P \rightarrow T_C$	20.90	0.7858	+2.64%	+1.95%
21-15	$T_P \rightarrow S_{P+1}$	21.39	0.7861	+3.58%	+1.00%

Table 2: BLEU & COMET scores on English-to-German OpenSubtitles of the EuroLLM 1.7B model with context size of 1 and the accuracy on the ContraPro dataset compared between modified and tuned head-relation pairs.

learned by the model, impacts the general translation performance and the contrastive accuracy. To this end, we replicate the Head Tuning technique from Mąka et al. (2025a), where the parameters of Q and K projections of the selected heads are tuned to attend more to the specified relation. We tune the heads of the EuroLLM 1.7B model with a context size of 1 using the ctxPro English-to-German Gender dataset. Next, we measure ContraPro accuracy and BLEU and COMET on the 10000 randomly selected English-to-German examples from the OpenSubtitles 2018 (Lison et al., 2018) dataset. The details are presented in Appendix D.

The results for the four selected head-relation pairs are presented in Table 2. The BLEU and COMET scores are stable compared to the base model, showing that when the change in attention scores is localized, the effect on general transla-

tion quality is minimal. The improvement in ContraPro accuracy can vary between head-relation pairs. While tuning the 19-4 $T_P \rightarrow T_{C+1}$ head-relation pair almost replicates the improvement of modifying, head tuning 21-15 $T_P \rightarrow S_{P+1}$ leads to an improvement of 1 percentage point compared to more than 3.5 when modified.

7 Indirect Object Identification Task

To test the generalizability of our method, we apply it to the task of Indirect Object Identification (IOI) (Wang et al., 2023a), where an expression such as "Then, [B] and [A] went to the house. [B] gave a drink to" (where "[A]" and "[B]" are names) should be completed with "[A]". We used EuroLLM 1.7B Instruct, but the task is very simple, and even this small model achieves near-perfect accuracy of 99.9%. Therefore, we follow the original paper and measure the difference in logits between the correct and incorrect name. The tested model exhibits a logit difference of 5.26. The details are presented in Appendix E.

We consider the token-to-token relations between the two names appearing in the examples. We mark the names in the first part of the IOI sequence as S_A and S_B for names "[A]" and "[B]", respectively, and as T_A and T_B in the second part. The results of modifying heads are presented in Figure 8, where the greatest improvement of more than 57% in the metric is achieved by 20-14 $T_A \rightarrow S_{B+1}$. Similar to the Context-aware MT setting, head-relation pairs that already show high average attention do not, in general, increase the metric when modified. Furthermore, we found heads that increase the metric when modified to attend different relations. For example, head 20-12 responds to $T_A \rightarrow S_{B+1}$, $T_A \rightarrow S_B$, $T_A \rightarrow T_{B+1}$, $T_A \rightarrow T_B$, and $T_A \rightarrow S_A$.

To test our method, we backpropagate the loss to the attention maps. The high accuracy on this task motivates the use of the log-probability Contrastive loss $\mathcal{L}_{logTC}$ (instead of the Max-Margin loss), which is directly related to the metric of the difference in the logits between "[A]" and "[B]". Using this loss with our method achieves Pearson and Spearman correlations of 0.6700 and 0.6862, respectively, and average precision of 0.6505. Furthermore, the top 3% head-relation pairs contain all results with a change in logit difference greater than 0.60 (top 37 head-relation pairs), while reducing the computational cost by almost 97%.

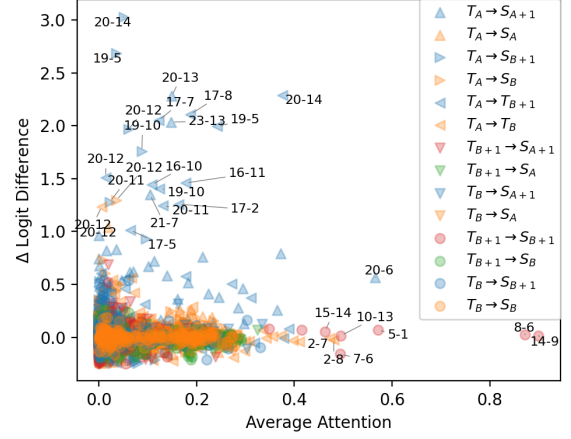


Figure 8: Average attention scores (x axis) and the change in the logit difference when modified to $C = 0.99$ for each head and relation on IOI dataset for EuroLLM 1.7B Instruct. For readability, we only show head-relation pairs with the positive change in logit difference.

This supports the generalizability and efficacy of our method.

8 Conclusions

In this work, we propose a gradient-based head attribution method based on the contrastive loss backpropagated to the attention maps. We exhaustively evaluate the method on the established tasks of disambiguating context-dependent phrases in Context-aware Machine Translation. We show that our method correlates with the Modifying Heads technique, allowing us a substantial expansion of the analysis scope, by reducing the computational cost by more than 96%. In total, we analyze 50 phenomena across 4 models and 4 language directions. We demonstrate the potential generalizability of our method by applying it to the task of Indirect Object Identification.

We identify that some heads are "general-purpose", improving the model's performance when attending to different relations. This behavior generalizes across different language directions and in some cases phenomena, and we link this overlap in the heads' apparent functions to the transfer effect between language directions and phenomena during training. Finally, our findings suggest that the models develop redundancies during training in terms of the head functions. In future work, we will expand the analysis to other contextual tasks and larger models.

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.
- Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 31–40.
- 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*.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.
- Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. [G-transformer for document-level 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 3442–3455, Online. Association for Computational Linguistics.
- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. [Identifying and controlling important neurons in neural machine translation](#). *arXiv preprint arXiv:1811.01157v1*.
- 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.
- Nikolay Bogoychev. 2021. [Not all parameters are born equal: Attention is mostly what you need](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 363–374, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations.
- Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091.
- Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.
- Linqing Chen, Junhui Li, Zhengxian Gong, Min Zhang, and Guodong Zhou. 2022. [One type context is not enough: Global context-aware neural machine translation](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(6).
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. [Efficiently exploring large language models for document-level machine translation with in-context learning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

- Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. [A survey of the state of explainable AI for natural language processing](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.
- Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. [Learn to remember: Transformer with recurrent memory for document-level machine translation](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costajussà. 2022. [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.
- Harritxu Gete, Thierry Etchegoyhen, and Gorka Labaka. 2023. [What works when in context-aware neural machine translation?](#) In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 147–156, Tampere, Finland. European Association for Machine Translation.
- Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. [Cross-attention is all you need: Adapting pretrained Transformers for machine translation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christian Hardmeier. 2012. [Discourse in statistical machine translation: A survey and a case study](#). *Discours-Revue de linguistique, psycholinguistique et informatique*, 11.
- Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. [Diving deep into context-aware neural machine translation](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. [Does neural machine translation benefit from larger context?](#) *arXiv preprint arXiv:1704.05135v1*.
- Jae-young Jo and Sung-Hyon Myaeng. 2020. [Roles and utilization of attention heads in transformer-based neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.
- Sehyeong Jo, Gangjae Jang, and Haesol Park. 2025. [Gmar: Gradient-driven multi-head attention rollout for vision transformer interpretability](#). In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 582–587.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. [Attention is not only a weight: Analyzing transformers with vector norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. [Incorporating Residual and Normalization Layers into Analysis of Masked Language Models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. 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, Steinhórfur 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.
- Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. [DecoderLens: Layerwise interpretation of encoder-decoder transformers](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4764–4780, Mexico City, Mexico. Association for Computational Linguistics.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.
- Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. [OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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.
- Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. [A simple and effective unified encoder for document-level machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. [Post-hoc interpretability for neural nlp: A survey](#). *ACM Comput. Surv.*, 55(8).
- Suvodeep Majumde, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. [A baseline revisited: Pushing the limits of multi-segment models for context-aware translation](#). *arXiv preprint arXiv:2210.10906v2*.
- Paweł Mała, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025a. [Analyzing the attention heads for pronoun disambiguation in context-aware machine translation models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6348–6377, Abu Dhabi, UAE. Association for Computational Linguistics.
- Paweł Mała, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025b. [You are what you train: Effects of data composition on training context-aware machine translation models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27414–27437, Suzhou, China. Association for Computational Linguistics.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.
- Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. [Selective attention for context-aware neural machine translation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2024. Locating and editing

- factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. [Document-level neural machine translation with hierarchical attention networks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. 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.
- Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. [Quantifying context mixing in transformers](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. 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.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, Tom Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. 2022. [In-context learning and induction heads](#). *arXiv preprint arXiv:2209.11895v1*.
- Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. [Salute the classic: Revisiting challenges of machine translation in the age of large language models](#). *Transactions of the Association for Computational Linguistics*, 13:73–95.
- 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.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Matt Post and Marcin Junczys-Dowmunt. 2023. [Escaping the sentence-level paradigm in machine translation](#). *arXiv preprint arXiv:2304.12959v1*.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115v2*.
- Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. [Eurollm-22b: Technical report](#). *arXiv preprint arXiv:2602.05879v1*.
- Ricardo Rei, Craig Stewart, Ana C 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.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. [Grad-cam: Visual explanations from deep networks via gradient-based localization](#). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. [Learning important features through propagating activation differences](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.
- Linxin Song, Yan Cui, Ao Luo, Freddy Lecue, and Irene Li. 2024. [Better explain transformers by illuminating important information](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2048–2062, St. Julian’s, Malta. Association for Computational Linguistics.
- Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. [Rethinking document-level neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 3319–3328. JMLR.org.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Ga l Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhan-shu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr s Gy rgy, Andr  Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrin, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim P der, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow,

- Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786v1*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. [BERT rediscovers the classical NLP pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Jörg Tiedemann and Yves Scherrer. 2017. [Neural machine translation with extended context](#). In *Proceedings of the Third Workshop on Dis-course in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Mariya Toneva and Leila Wehbe. 2019. *Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)*. Curran Associates Inc., Red Hook, NY, USA.
- Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. [Context gates for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 5:87–99.
- 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.
- Jesse Vig and Yonatan Belinkov. 2019. [Analyzing the structure of attention in a transformer language model](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. [Context-aware monolingual repair for neural machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. [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.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019c. [Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.
- Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. [Gradient-based analysis of NLP models is manipulable](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations*.
- Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui,

- Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. [Benchmarking and improving long-text translation with large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- Rachel Wicks and Matt Post. 2023. [Identifying context-dependent translations for evaluation set production](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- Billy T. M. Wong and Chunyu Kit. 2012. [Extending machine translation evaluation metrics with lexical cohesion to document level](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.
- 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:2401.06468v4*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671v4*.
- Kayo Yin and Graham Neubig. 2022. [Interpreting language models with contrastive explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. [Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.
- Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. [The same but different: Structural similarities and differences in multilingual language modeling](#). In *The Thirteenth International Conference on Learning Representations*.
- Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2021. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3983–3989.

A Generating Contrastive Dataset

In this section, we describe the creation process of the contrastive version of the ctxPro datasets (Wicks and Post, 2023). Below, we list the contextually-dependent linguistic phenomena that we included in our study:

- **Gender** (anaphoric pronouns) - translating a pronoun from a non-gendered language to a language with gendered nouns.
- **Formality** (anaphoric pronouns) - translating into a language with different second-person pronouns distinguishing intimate from formal relationships between speakers from a language lacking this distinction.
- **Auxiliary** (verb phrase ellipsis) - translating into a language that require the head of the verb phrase from a language that allows for only the modal or auxiliary to be used.
- **Inflection** (verb phrase ellipsis) - translating into a language with noun morphology dependent on the grammatical role from a language where this is not the case.

We started with the evaluation datasets released with the paper under <https://github.com/rewicks/ctxpro/tree/main/paper/jsons>. For all phenomena and language directions, we replaced the contextually dependent phrase in the target segment with phrases sampled from the list of phrases from the other examples (excluding the same phrases), respecting the capitalization of the original phrase. We sampled with respect to the distribution of occurrences, as our tests showed that this increased the difficulty of the created datasets. We chose the number of contrastive target sentences for each example to balance the difficulty (based on the EuroLLM 1.7B model) and the computational cost of evaluation. We set this number to 4 for gender, 5 for formality and inflection, and 8 for auxiliary. In the experiments, we used ContraPro (Müller et al., 2018) and LCPT (Lopes et al., 2020) instead of the generated contrastive versions of ctxPro for Gender in English-to-German and English-to-French, respectively.

B Evaluation Details

We evaluated the models on a single GPU (NVIDIA GeForce RTX 3090 24GB for EuroLLM 1.7B Context 1, Qwen 2.5 1.5B, and Gemma 3 1B, and NVIDIA H100 80GB for EuroLLM 1.7B Context 5 and EuroLLM 9B). We used 5 previous sentences as context on both the source and target side (apart from the experiment with EuroLLM 1.7B Context 1). We used a batch size of 2 for evaluation and 1 for backpropagating gradients. The models were loaded with precision bf16 (all models from the EuroLLM family) and 8-bit (Qwen 2.5 and Gemma 3). We set the maximum length of the prompt to 2048 for EuroLLM 1.7B and Gemma 3 1B, and 4096 for EuroLLM 9B and Qwen 2.5 1.5B.

To prepare the inputs to the models, we concatenated the context and current sentences on both the source and target sides. We tokenized the prompt shown in Listing 1 using the models’ tokenizers and passed it through the models to obtain the token probabilities. We used the system prompts found in the model cards on HuggingFace. The sizes of the used datasets are presented in Table 3.

Language Direction	Phenomenon	Examples	Contrastive
English-to-German	ContraPro	12000	3.0
	Gender	13640	5.0
	Auxiliary	3180	9.0
	Formality	10800	6.0
English-to-French	LCPT	14000	2.0
	Gender	19098	5.0
	Auxiliary	8322	9.0
	Formality	12000	5.0
English-to-Polish	Gender	14855	5.0
	Auxiliary	8085	9.0
	Formality	15216	6.0
	Inflection	12000	6.0
English-to-Russian	Gender	10305	5.0
	Auxiliary	8667	9.0
	Formality	10075	6.0
	Inflection	12000	6.0

Table 3: The sizes of the used datasets in terms of the number of examples and the average number of contrastive targets (including the correct target) for each dataset and language direction.

Listing 1: Input template used for evaluating the models. The concatenated context and current sentences on the source and target sides are represented as `{src}` and `{tgt}` respectively. The source and target languages are specified in natural language (e.g., "English", "German") and marked as `{src_lang}` and `{tgt_lang}` respectively. We used the system prompts found in the model cards on HuggingFace.

```
{
  "role": "system",
  "content": {system_prompt}
},
{
  "role": "user",
  "content": "Translate the following {src_lang} text to {tgt_lang}:
    ↪ '{src}' "
},
{
  "role": "assistant",
  "content": {tgt}
}
```

C Time and Memory Requirements

The Modifying Heads technique requires repeated evaluations for intervening in attention for each head-relation pair. Table 4 shows the number of layers, heads, and evaluations (assuming 10 token-to-token relations that we used in this study) for each tested model. The largest of our models (EuroLLM 9B Instruct) requires 133400 separate evaluations.

In Tables 5 and 6, we present the average execution time and peak GPU memory⁶ for every model, language direction, and phenomenon. The calculation of gradients used more GPU memory compared to the evaluation, even with the lower batch size. Although the execution time is also increased, this increase is small in comparison to the total time required for the full sweep over the head-relation pairs, because of the very high number of evaluations. For example, assuming evaluating 3% of the top head-

⁶We used the `torch.cuda.memory.max_memory_allocated()` function.

relation pairs, the total computation time for the EuroLLM 9B Instruct model on ContraPro is 175 GPU hours. Meanwhile, the full sweep would take 5824 GPU hours (our method requires 3% of the full sweep time). The proportion of time spent on calculating gradients is smaller for models containing more heads. For Gemma 3 1B with only 104 heads in the model, the full sweep of ContraPro requires 537 GPU hours compared to 21 hours for our method (4% of the full sweep).

Model	Layers	Heads	Heads Total	Evaluations	Our Method
EuroLLM 1.7B Instruct	24	16	384	3840	120
EuroLLM 9B Instruct	42	32	1344	13440	400
Qwen 2.5 1.5B Instruct	28	12	336	3360	100
Gemma 3 1B Instruct	24	4	104	1040	40

Table 4: The number of layers and heads, the total number of heads, the number of evaluations for each language direction and phenomenon (assuming 10 relevant token-to-token relations), and the number of evaluations when applying our method for each tested model.

Model	Language Direction	Gender		Auxiliary		Formality		Inflection	
		Time	Mem.	Time	Mem.	Time	Mem.	Time	Mem.
EuroLLM 1.7B Instruct [†]	en-de	0:12	9.0	0:08	13.8	0:12	11.3	-	-
	en-fr	0:12	7.8	0:22	13.0	0:11	10.1	-	-
	en-pl	0:24	13.2	0:21	14.1	0:17	12.6	0:13	9.9
	en-ru	0:16	14.0	0:23	13.2	0:11	12.1	0:13	10.7
EuroLLM 9B Instruct [†]	en-de	0:26	25.7	0:17	31.8	0:27	28.5	-	-
	en-fr	0:25	24.2	0:48	30.6	0:26	27.0	-	-
	en-pl	0:52	32.1	0:46	32.3	0:40	30.7	0:26	26.3
	en-ru	0:36	33.7	0:49	30.9	0:26	29.9	0:27	27.7
Qwen 2.5 1.5B Instruct [*]	en-de	0:35	9.1	0:20	14.4	0:35	12.1	-	-
	en-fr	0:35	7.1	0:54	13.1	0:34	10.2	-	-
	en-pl	1:06	15.4	0:58	16.0	0:53	15.3	0:39	11.1
	en-ru	0:45	11.9	1:00	14.5	0:35	12.9	0:39	11.5
Gemma 3 1B Instruct [*]	en-de	0:30	11.3	0:17	19.9	0:30	15.4	-	-
	en-fr	0:33	8.8	0:45	18.3	0:31	13.4	-	-
	en-pl	1:06	19.6	0:40	23.2	0:43	18.6	0:31	13.8
	en-ru	0:34	16.6	0:42	19.5	0:27	16.3	0:31	14.3

Table 5: The execution time (in hours:minutes) and peak memory (in gigabytes) of a single evaluation for each model and phenomenon. [†]NVIDIA H100 80GB. ^{*}NVIDIA GeForce RTX 3090 24GB.

Model	Language Direction	Gender		Auxiliary		Formality		Inflection	
		Time	Mem.	Time	Mem.	Time	Mem.	Time	Mem.
EuroLLM 1.7B Instruct [†]	en-de	0:19	15.9	0:06	21.5	0:17	17.9	-	-
	en-fr	0:23	12.9	0:15	20.5	0:19	16.3	-	-
	en-pl	0:25	20.7	0:15	22.0	0:25	19.9	0:19	16.0
	en-ru	0:18	21.9	0:16	20.7	0:16	19.1	0:18	17.1
EuroLLM 9B Instruct [†]	en-de	0:39	58.3	0:16	72.0	0:40	62.2	-	-
	en-fr	0:43	51.1	0:42	69.4	0:41	59.2	-	-
	en-pl	0:58	70.4	0:39	73.2	0:58	68.1	0:40	58.1
	en-ru	0:40	73.7	0:42	69.9	0:37	66.2	0:41	61.3
Qwen 2.5 1.5B Instruct [*]	en-de	0:55	13.4	0:19	18.8	0:55	15.8	-	-
	en-fr	1:00	9.3	0:50	17.0	0:56	13.2	-	-
	en-pl	1:19	20.2	0:51	21.0	1:20	20.1	0:58	14.5
	en-ru	0:54	15.6	0:53	18.9	0:51	16.9	0:58	15.1
Gemma 3 1B Instruct [*]	en-de	0:59	13.3	0:17	20.3	0:30	15.4	-	-
	en-fr	1:08	9.4	0:45	18.8	0:58	13.9	-	-
	en-pl	1:15	20.3	0:44	23.2	1:16	19.1	0:58	13.4
	en-ru	0:51	16.1	0:49	20.0	0:49	16.9	0:57	14.8

Table 6: The execution time (in hours:minutes) and peak memory (in gigabytes) of a single gradient calculation run for each model and phenomenon. [†]NVIDIA H100 80GB. ^{*}NVIDIA GeForce RTX 3090 24GB.

D Details of Head Tuning

We perform Head Tuning using the ctxPro English-to-German Gender dataset containing 13640 annotated examples. We freeze all parameters of the model apart from the Q and K projections of the selected layer, and calculate the MSE loss between the actual attention scores of the selected head and the attention scores after applying the intervention. Following Mąka et al. (2025a), we use the baseline model to calculate the modified attention scores (we load the base model twice: the first one for fine-tuning, the second one for calculating the target attention scores). We used a learning rate of $2e-5$ with linear scheduling and no warmup steps, 10 epochs, and a batch size of 4 with 8 gradient accumulation steps. The training took less than one hour on NVIDIA GeForce RTX 3090 24GB.

We evaluate the models on ContraPro contrastive dataset and on 10,000 English-to-German examples sampled from OpenSubtitles 2018 dataset (Lison et al., 2018). The translation used the greedy decoding (the beam size of 1 and no sampling) with the batch size of 16. The context size was set to 1 and we used the gold (provided) target-side context. We measured BLEU (Papineni et al., 2002) using the sacreBLEU implementation⁷ (Post, 2018) and COMET⁸ (Rei et al., 2020).

E Details of Indirect Object Identification Experiment

In the task of Indirect Object Identification (IOI) (Wang et al., 2023a) the model has to complete expression in the form "Then, [B] and [A] went to the house. [B] gave a drink to ", where "[A]" and "[B]" are names, and the model should prefer "[A]" over "[B]". In the actual examples, both name placeholders are replaced by popular English names. In the original paper, the names are chosen to be encoded as a single token by the tested model. In case of our model (EuroLLM 1.7B Instruct), the list contained multi-token names, therefore, we removed them when generating the examples. The above example is in the form BABA (corresponding to the order of names). In our experiments we include both BABA and ABBA templates. We generated 10000 unique examples using the code from the paper’s repository⁹. Due to the simplicity of the task and the near-perfect accuracy exhibited by the tested model, we measure the difference in the logits of the two names from the examples.

We evaluate the model on NVIDIA GeForce RTX 3090 24GB. A single evaluation takes 8 minutes and the registered peak memory on the GPU is 3.1 gigabyte. Calculating gradients lasts 19 minutes with the peak memory of 6.8 gigabytes. To define the relations of interest, we mark the names in the first part of the IOI sequence as S_A and S_B for names "[A]" and "[B]", respectively, and as T_A and T_B in the second part. We consider the following 14 relations:

- $T_A \rightarrow S_A$,
- $T_A \rightarrow S_{A+1}$,
- $T_A \rightarrow S_B$,
- $T_A \rightarrow S_{B+1}$,
- $T_A \rightarrow T_B$,
- $T_A \rightarrow T_{B+1}$,
- $T_B \rightarrow S_A$,
- $T_B \rightarrow S_{A+1}$,
- $T_B \rightarrow S_B$,
- $T_B \rightarrow S_{B+1}$,
- $T_{B+1} \rightarrow S_A$,
- $T_{B+1} \rightarrow S_{A+1}$,
- $T_{B+1} \rightarrow S_B$,
- $T_{B+1} \rightarrow S_{B+1}$.

Consequently, we perform 53,760 evaluations modifying heads to $C = 0.99$ targeting each head-relation pair. This amounts to 7168 GPU hours, compared to 216 hours for our method when evaluating top 3% of head-relation pairs, reducing the total time to 3.01%.

⁷We used <https://github.com/mjpost/sacrebleu> with default parameters.

⁸We used the Unbabel/wmt22-comet-da model.

⁹<https://github.com/redwoodresearch/Easy-Transformer>

F Extended Results

Model	Language Direction	Gender	Auxiliary	Formality	Inflection
EuroLLM 1.7B Instruct	en-de	84.8 ^{+0.8} _{-0.8}	88.0 ^{+1.4} _{-1.5}	88.9 ^{+0.8} _{-0.8}	-
	en-fr	94.6 ^{+0.5} _{-0.5}	91.3 ^{+0.8} _{-0.8}	91.0 ^{+0.6} _{-0.7}	-
	en-pl	95.4 ^{+0.4} _{-0.5}	87.3 ^{+0.9} _{-0.9}	70.2 ^{+1.0} _{-0.6}	94.0 ^{+0.5} _{-0.6}
	en-ru	90.9 ^{+0.7} _{-0.7}	89.2 ^{+0.8} _{-0.9}	84.0 ^{+0.9} _{-1.0}	96.2 ^{+0.4} _{-0.5}
	en-de	87.8 ^{+0.8} _{-0.8}	58.1 ^{+2.2} _{-2.2}	90.9 ^{+0.7} _{-0.7}	-
EuroLLM 9B Instruct	en-fr	95.8 ^{+0.4} _{-0.4}	69.0 ^{+1.3} _{-1.3}	65.9 ^{+1.1} _{-1.1}	-
	en-pl	94.8 ^{+0.5} _{-0.5}	63.9 ^{+1.3} _{-1.4}	56.1 ^{+1.0} _{-1.0}	81.5 ^{+0.9} _{-0.9}
	en-ru	78.1 ^{+1.1} _{-1.1}	63.7 ^{+1.4} _{-1.4}	61.4 ^{+1.2} _{-1.3}	85.8 ^{+0.8} _{-0.9}
	en-de	69.9 ^{+1.1} _{-1.1}	86.4 ^{+1.5} _{-1.6}	85.9 ^{+0.9} _{-0.8}	-
Qwen 2.5 1.5B Instruct	en-fr	91.5 ^{+0.6} _{-0.6}	89.1 ^{+0.9} _{-0.9}	91.3 ^{+0.6} _{-0.7}	-
	en-pl	81.8 ^{+0.8} _{-0.8}	74.5 ^{+1.2} _{-1.2}	53.4 ^{+1.0} _{-1.0}	87.9 ^{+0.8} _{-0.8}
	en-ru	85.5 ^{+0.9} _{-1.0}	82.5 ^{+1.0} _{-1.1}	85.8 ^{+0.9} _{-0.9}	95.3 ^{+0.5} _{-0.5}
	en-de	73.4 ^{+1.0} _{-1.0}	57.7 ^{+2.2} _{-2.2}	79.9 ^{+1.0} _{-1.0}	-
Gemma 3 1B Instruct	en-fr	89.8 ^{+0.7} _{-0.7}	71.3 ^{+1.3} _{-1.3}	79.2 ^{+0.9} _{-1.0}	-
	en-pl	82.3 ^{+0.8} _{-0.8}	50.9 ^{+1.5} _{-1.4}	45.4 ^{+1.0} _{-1.1}	80.2 ^{+0.9} _{-1.0}
	en-ru	76.1 ^{+1.1} _{-1.1}	56.4 ^{+1.4} _{-1.3}	78.2 ^{+1.1} _{-1.0}	88.1 ^{+0.8} _{-0.7}
	en-de	84.8 ^{+0.8} _{-0.8}	88.0 ^{+1.4} _{-1.5}	88.9 ^{+0.8} _{-0.8}	-

Table 7: Results in terms of accuracy on the contrastive datasets for different language directions and phenomena of the unmodified models. Superscript and subscript show the 99% confidence intervals.

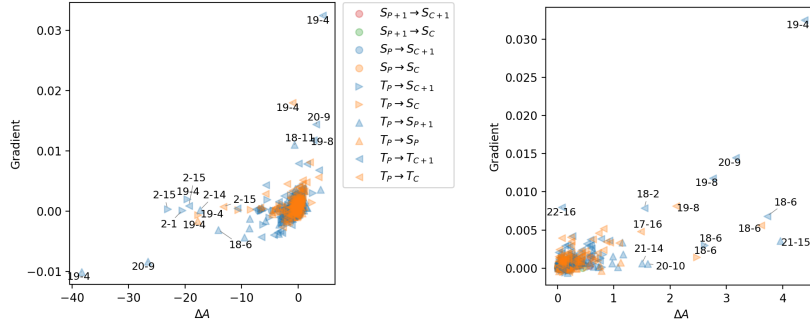


Figure 9: The relation between the accumulated gradients based on the Token-level Max-Margin Loss $\mathcal{L}_{TMM}$ and the change in the accuracy ΔA when modified to $C = 0.99$ for all heads and different relations for EuroLLM 1.7B Instruct on ContraPro with context size of 5. Left figure shows all results, while the right figure shows only the head-relation pairs for the values of $\Delta A > 0$.

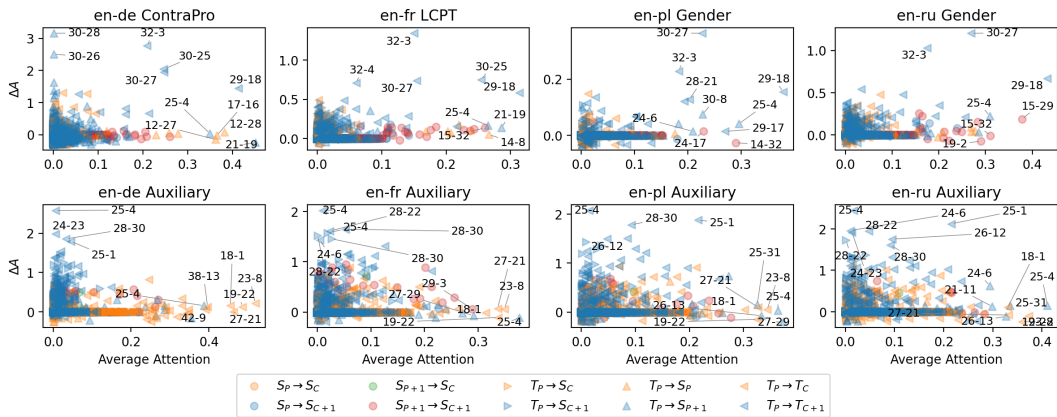


Figure 10: The results in terms of average attention scores (x axes) and accuracy increase ΔA (y axes) for head-relation pairs for **EuroLLM 9B Instruct** on all tested phenomena and language directions. To improve clarity, we omitted the head-relation pairs with a large negative ΔA .

Model	Setting	Loss	All		$\Delta A > 0$		$\Delta A \leq 0$	
			r	ρ	r	ρ	r	ρ
EuroLLM 1.7B	ContraPro Context 5	Attention	0.0459	0.1756	0.1565	-0.1416	0.0187	0.2322
		$\mathcal{L}_{NLL}$	-0.0364	-0.2177	0.1677	0.1659	-0.0357	-0.2502
		$\mathcal{L}_{logC}$	0.0816	0.1217	0.5740	0.2300	0.0219	0.0855
		$\mathcal{L}_{logTC}$	0.0782	0.0363	0.6267	0.3796	-0.0140	-0.0764
		$\mathcal{L}_{TC}$	0.2181	0.1178	0.6492	0.4480	0.1760	-0.0049
		$\mathcal{L}_{logMM}$	0.2378	0.2008	0.6544	0.4060	0.2156	0.1232
		$\mathcal{L}_{logTMM}$	0.2324	0.1469	0.6747	0.4972	0.2118	0.0183
EuroLLM 1.7B	ContraPro Context 1	$\mathcal{L}_{TMM}$	0.2768	0.1711	0.7028	0.5268	0.2737	0.0423
		Attention	0.0329	0.0650	0.1368	-0.1433	0.0244	0.1241
		$\mathcal{L}_{NLL}$	-0.0086	-0.1579	0.0722	0.1342	-0.0171	-0.2014
		$\mathcal{L}_{logC}$	0.2221	0.2363	0.6180	0.3536	0.1972	0.1640
		$\mathcal{L}_{logTC}$	0.3192	0.2869	0.6621	0.5059	0.3379	0.1494
		$\mathcal{L}_{TC}$	0.3017	0.2902	0.6509	0.4798	0.3041	0.1547
		$\mathcal{L}_{logMM}$	0.2730	0.2284	0.6194	0.4043	0.2738	0.1382
EuroLLM 9B*	ContraPro	$\mathcal{L}_{logTMM}$	0.2974	0.2790	0.6451	0.5151	0.3272	0.1366
		$\mathcal{L}_{TMM}$	0.3492	0.3330	0.6217	0.5463	0.4033	0.1924
		Attention	0.0248	-0.0954	0.3834	0.0298	-0.0405	-0.1075
		$\mathcal{L}_{NLL}$	-0.0432	-0.2059	0.0254	-0.0005	-0.0183	-0.1865
		$\mathcal{L}_{logC}$	0.1904	0.0380	0.4259	0.1751	0.1827	-0.0020
		$\mathcal{L}_{logTC}$	0.2308	0.1680	0.4639	0.2495	0.2383	0.0847
		$\mathcal{L}_{TC}$	0.2817	0.1927	0.4689	0.2550	0.3124	0.1142
Qwen 1.5B	ContraPro	$\mathcal{L}_{logMM}$	0.3597	0.1671	0.6416	0.2758	0.4185	0.0807
		$\mathcal{L}_{logTMM}$	0.3403	0.2316	0.6278	0.3064	0.3975	0.1248
		$\mathcal{L}_{TMM}$	0.4234	0.3142	0.7054	0.3618	0.4897	0.1921
		Attention	0.0528	0.0783	0.0996	-0.1318	0.0480	0.0737
		$\mathcal{L}_{NLL}$	-0.0602	-0.0897	0.1302	0.1592	-0.0785	-0.1658
		$\mathcal{L}_{logC}$	0.0970	0.2573	0.6234	0.3410	0.0419	0.1098
		$\mathcal{L}_{logTC}$	0.1148	0.2395	0.7061	0.4574	0.0686	0.0505
Qwen 1.5B	LCPT	$\mathcal{L}_{TC}$	0.1225	0.2311	0.6950	0.4366	0.0834	0.0504
		$\mathcal{L}_{logMM}$	0.1111	0.2888	0.7113	0.4125	0.0549	0.1165
		$\mathcal{L}_{logTMM}$	0.1214	0.2664	0.7333	0.4855	0.0754	0.0654
		$\mathcal{L}_{TMM}$	0.1332	0.3108	0.7470	0.4939	0.0913	0.1139
		Attention	0.0496	-0.0205	0.1790	-0.0810	0.0685	0.0906
		$\mathcal{L}_{NLL}$	-0.0480	0.0527	0.1578	0.2352	-0.0704	-0.2398
		$\mathcal{L}_{logC}$	0.0617	0.2465	0.7668	0.2906	0.0265	0.1081
Qwen 1.5B	Auxiliary	$\mathcal{L}_{logTC}$	0.0708	0.2856	0.7743	0.3356	0.0449	0.0973
		$\mathcal{L}_{TC}$	0.0600	0.2806	0.8038	0.3811	0.0118	0.0577
		$\mathcal{L}_{logMM}$	0.0469	0.2144	0.8239	0.3887	-0.0067	-0.0255
		$\mathcal{L}_{logTMM}$	0.0575	0.2319	0.8326	0.4271	0.0171	-0.0441
		$\mathcal{L}_{TMM}$	0.0459	0.2826	0.8307	0.3587	-0.0187	0.0737
		Attention	-0.0979	0.0491	0.0072	0.0638	-0.1230	-0.0053
		$\mathcal{L}_{NLL}$	-0.0303	0.0386	0.6507	0.4656	-0.0885	-0.2675
Qwen 1.5B	ContraPro	$\mathcal{L}_{logC}$	0.0687	0.2875	0.8459	0.5902	0.0148	0.0312
		$\mathcal{L}_{logTC}$	0.0718	0.2449	0.8707	0.6149	0.0248	-0.0478
		$\mathcal{L}_{TC}$	0.0749	0.2366	0.8724	0.5907	0.0286	-0.0491
		$\mathcal{L}_{logMM}$	0.0609	0.2445	0.8023	0.5688	0.0076	-0.0128
		$\mathcal{L}_{logTMM}$	0.0697	0.2238	0.8381	0.6068	0.0280	-0.0696
		$\mathcal{L}_{TMM}$	0.0820	0.2430	0.8803	0.5984	0.0432	-0.0450
		Attention	0.0502	0.0908	0.1447	-0.0356	0.0006	0.0293
Gemma 1B	ContraPro	$\mathcal{L}_{NLL}$	-0.1424	-0.2230	0.2575	0.3173	-0.1423	-0.2481
		$\mathcal{L}_{logC}$	0.2814	0.2022	0.5688	0.3447	0.2779	0.1632
		$\mathcal{L}_{logTC}$	0.3139	0.2058	0.6048	0.4330	0.3464	0.1534
		$\mathcal{L}_{TC}$	0.3886	0.2508	0.6477	0.4579	0.4340	0.1927
		$\mathcal{L}_{logMM}$	0.3336	0.2274	0.6206	0.3762	0.3652	0.1765
		$\mathcal{L}_{logTMM}$	0.3375	0.2366	0.6271	0.4376	0.3778	0.1746
		$\mathcal{L}_{TMM}$	0.4155	0.2892	0.6712	0.4698	0.4613	0.2113

Table 8: Pearson r and Spearman ρ correlation coefficients between accumulated gradients and difference in accuracy ΔA for different sets of head-relation pairs (all, $\Delta A > 0$, and $\Delta A \leq 0$) when attention scores are modified to $C = 0.99$ for all relations of interest for all models and experimental settings (we also include the results when using average attention scores instead of accumulated gradients). *EuroLLM 9B was evaluated on the limited set of token-to-token relations.

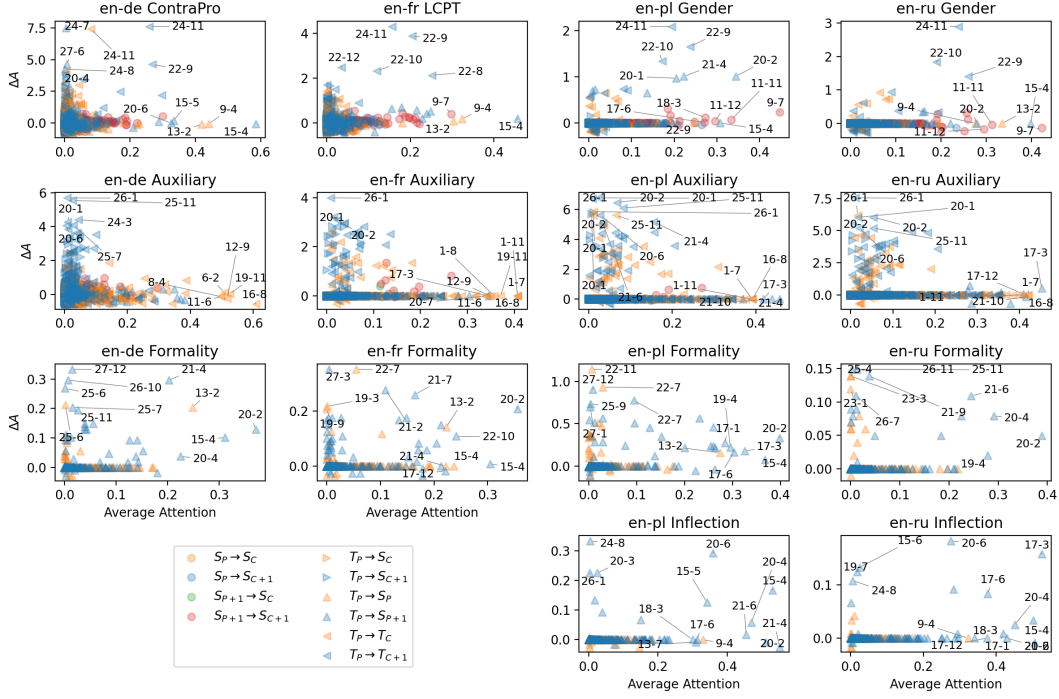


Figure 11: The results in terms of average attention scores (x axes) and accuracy increase ΔA (y axes) for head-relation pairs for **Qwen 2.5 1.5B Instruct** on all tested phenomena and language directions. To improve clarity, we omitted the head-relation pairs with a large negative ΔA .

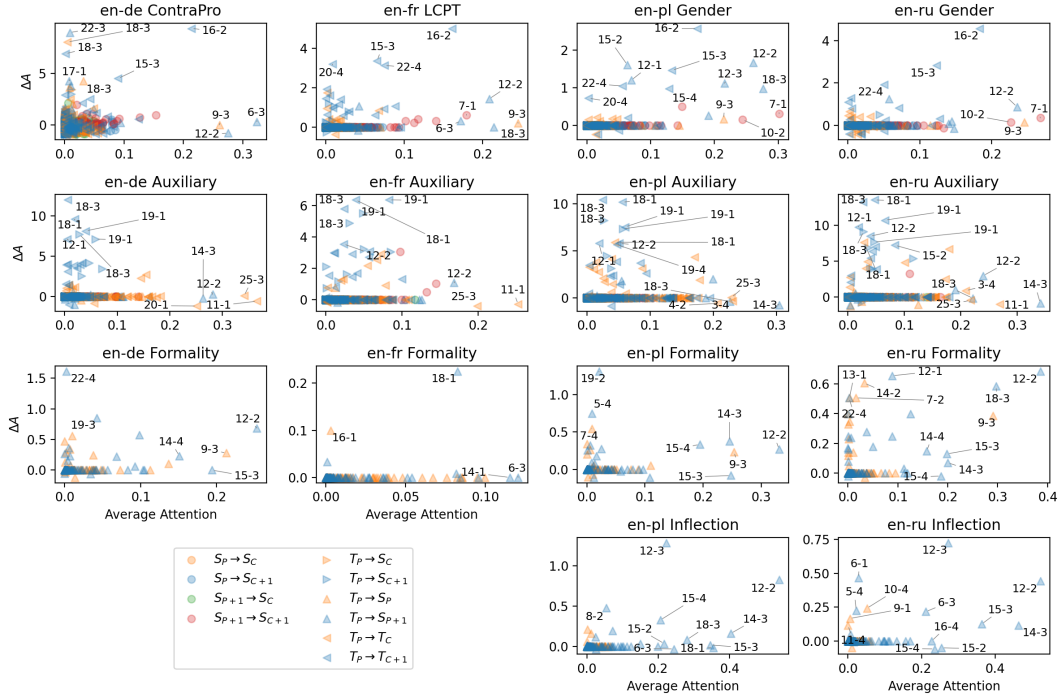


Figure 12: The results in terms of average attention scores (x axes) and accuracy increase ΔA (y axes) for head-relation pairs for **Gemma 3 1B Instruct** on all tested phenomena and language directions. To improve clarity, we omitted the head-relation pairs with a large negative ΔA .