A Representation Level Analysis of NMT Model Robustness to Grammatical Errors

Abderrahmane Issam
Yusuf Can Semerci
Jan Scholtes
Gerasimos Spanakis
Department of Advanced Computing Sciences
Maastricht University
{abderrahmane.issam, y.semerci, j.scholtes, jerry.spanakis}@maastrichtuniversity.nl

Abstract

Understanding robustness is essential for building reliable NLP systems. Unfortunately, in the context of machine translation, previous work mainly focused on documenting robustness failures or improving robustness. In contrast, we study robustness from a model representation perspective by looking at internal model representations of ungrammatical inputs and how they evolve through model layers. For this purpose, we perform Grammatical Error Detection (GED) probing and representational similarity analysis. Our findings indicate that the encoder first detects the grammatical error, then corrects it by moving its representation toward the correct form. To understand what contributes to this process, we turn to the attention mechanism where we identify what we term *Robustness Heads*. We find that *Robustness Heads* attend to interpretable linguistic units when responding to grammatical errors, and that when we fine-tune models for robustness, they tend to rely more on *Robustness Heads* for updating the ungrammatical word representation. [1]

1 Introduction

Neural Machine Translation (NMT) has seen great success, especially since the introduction of the Transformer architecture (Vaswani et al., 2017). Recent advances in NMT introduced models that can translate between over 200 languages (NLLB Team et al., 2022). While this achievement is impressive and drives the deployment in real-world scenarios, evaluating and understanding NMT robustness remains essential for building reliable NMT systems.

Early works have focused on documenting the robustness failures of NMT models, or improving their robustness (Napoles et al., 2016; Khayrallah and Koehn, 2018; Belinkov and Bisk, 2018; Anastasopoulos, 2019; Jayanthi and Pratapa, 2021). However, there has been limited analysis of model representations in response to noise. Therefore, our goal in this work is to fill this gap by analyzing robustness from a representation perspective.

Our hypothesis is that the encoder detects and corrects the representation of the ungrammatical word by moving its representation toward the correct form. To study the detection part, we use GED probing to evaluate how the accuracy of detecting the ungrammatical word changes through the encoder layers. For the correction part, we measure the representation distance between the ungrammatical word and its correct grammatical form. We find that generally the probing performance increases in the first half layers of the model, then plateaus or decreases, while the representation distance on the overall decreases along model layers.

To understand what contributes to correcting the representations, we turn to the attention mechanism due to its crucial contribution to transformer model performance. We identify what we term *Robustness Heads*, which are attention heads that contribute to moving the ungrammatical word's representation toward its correct form. We find that after fine-tuning models on ungrammatical sentences, and thus, making them more robust, they learn to rely more on *Robustness Heads* for updating the ungrammatical word's representation especially in deeper layers where we hypothesize that the correction is happening.

Our work addresses the following research questions and makes the following contributions:
RQ1: *How do models represent and handle grammatical errors to achieve robustness?* NMT encoders inherently implement a Grammatical Error Correction (GEC) setup in which they first detect the ungrammatical word, then correct it by moving its representation toward the grammatical form.
RQ2: *How does grammatical error representation differ across models and languages?* Models

[1] Our code: https://github.com/issam9/nmt-robustness-analysis.

Findings of the Association for Computational Linguistics: ACL 2025, pages 8579–8601
July 27 - August 1, 2025

respond similarly to grammatical errors however we find differences across languages which we attribute to their linguistic differences.

RQ3: *How does fine-tuning for robustness influences models to achieve improved robustness?* Fine-tuned models exhibit similar behavior to their base models but tend to rely more on *Robustness Heads* for handling grammatical errors and enhancing their resilience to noisy inputs.

2 Related Work

There have been multiple works analyzing transformer models representations for different purposes. Probing has been explored for understanding information that is encoded in pre-trained model representations (Belinkov et al., 2017; Ettinger, 2020; Belinkov et al., 2020; Liu et al., 2021; Davis et al., 2022), as well as to understand the effect of fine-tuning (Mosbach et al., 2020; Durrani et al., 2021; Zhou and Srikumar, 2022; Durrani et al., 2023a), and contextual embeddings (Tenney et al., 2019; Klafka and Ettinger, 2020). The attention module, given its importance in the success of the transformer architecture, has also been explored for interpreting transformer models (Raganato and Tiedemann, 2018; Clark et al., 2019; Voita et al., 2019; Zhang et al., 2023). Other works have focused on understanding and comparing between different models through the lens of similarity analysis (Kudugunta et al., 2019; Wu et al., 2020; Vázquez et al., 2021). In our work, we combine all these techniques to understand NMT models robustness to grammatical errors.

Fine-tuning on downstream tasks leads to changes in representations that might not be favorable (e.g. catastrophic forgetting), and thus has been an active area of analysis (Merchant et al., 2020; Durrani et al., 2021; Phang et al., 2021; Zhou and Srikumar, 2022; Durrani et al., 2023b; Neerudu et al., 2023). Our work focuses on understanding the effects of fine-tuning for robustness to grammatical errors in the context of NMT, where we look at both model representations and downstream NMT performance.

Robustness is of critical importance for NLP models. Early works have explored improving robustness to different types (e.g. User Generated Data, Automatic Speech Recognition, Non native speakers, ...) of noise by using synthetic data (Anastasopoulos et al., 2019; Zhou et al., 2019; Karpukhin et al., 2019; Salesky et al., 2019; Jayanthi and Pratapa, 2021; Wang et al., 2021; Zhao and Calapodescu, 2022). Recent works show that even Large Language Models (LLMs) witness performance degradation when confronted with synthetic noise (Chen et al., 2024; Zhu et al., 2024). Multiple techniques were introduced to deal with noise or adversarial attacks by pushing noisy or adversarial samples representations to be similar to those of the original samples (Xu et al., 2021; Passban et al., 2021; Yang et al., 2022; Wang et al., 2023) and our experiments show that NMT encoders do this inherently when trained on synthetic noise. While analyzing the robustness of NMT transformer models has been an area of exploration (Napoles et al., 2016; Khayrallah and Koehn, 2018; Belinkov and Bisk, 2018; Anastasopoulos, 2019; Jayanthi and Pratapa, 2021), the focus was more on documenting model failures under noise or adversarial attacks rather than analyzing the model internals, therefore, our work is an attempt to fill this gap. In computer vision, however, Cianfarani et al. (2022) explored representation similarity to understand adversarially trained Deep Neural Networks (DNNs) and compare them to non-robust DNNs, while we focus on understanding the effect of fine-tuning NMT transformer models on synthetic grammatical errors, and we compare them against their base model as a less robust model, as well as against a domain adapted model. In addition, we probe the linguistic features encoded in the representations and we compare the models on the basis of their attention mechanism.

3 Methodology

Our analysis starts by introducing grammatical errors in the dataset, which we describe in §3.1. Subsequently, in §3.2 we describe how we fine-tune models for robustness to analyze the effects of fine-tuning. Finally, in §3.3 we present the methods of our analysis, namely: GED probing, representation similarity, and *Robustness Heads*, where we describe our method for finding *Robustness Heads*, and analyzing their attention to POS tags.

3.1 Synthetic Grammatical Errors

To provide a representation level analysis of robustness, it is crucial to have granular control over grammatical errors. We achieve this by introducing synthetic grammatical errors into clean sentences. We focus on three types of grammatical errors that are common in non-native speaker language (Izumi

et al., 2004; Napoles et al., 2016; Anastasopoulos, 2019), and we create an adversarial copy of the dataset for each type, where we insert one error per sentence when possible. We focus on grammatical errors with clear linguistic functions to be able to link our analysis to the linguistic features of the source language. However, to validate the generalizability of our analysis to different types of errors as well as more than one error per sentence, we use MORPHEUS (Tan et al., 2020) as a black-box adversarial attack that greedily introduces inflectional errors to minimize the BLEU score.

We follow the implementation of (Anastasopoulos et al., 2019) to introduce article (*Article*), preposition (*Prep*) and noun number (*Nounnum*) replacement errors in the dataset. For each sentence $X = \{w_1, w_2, w_3, ..., w_n\}$ in the dataset D, we introduce one of the three grammatical errors in X when possible to get $\widetilde{X} = \{w_1, w_2, \widetilde{w_3}, ..., w_n\}$, where $\widetilde{w_3}$ is the noisy word that was sampled to replace w_3. Therefore, $\widetilde{w_3}$ represents the ungrammatical word, and w_3 is its target grammatical form. The result is an adversarial dataset $\widetilde{D}$ for each error type.

3.2 Fine-tuning for Robustness

Our analysis focuses on four well established NMT models, namely: OPUS-MT, M2M100, MBART and NLLB. We fine-tune these models on the adversarial dataset of one of the error types, and since this leads to improving their robustness to the error type, we also analyze the representations of fine-tuned models. To separate the effect of robustness from domain adaptation, we compare against a version of the model that is fine-tuned on the clean version of the data. We refer to these models as Base, Noise-Finetuned, and Clean-Finetuned respectively. However, we only fine-tune the encoder given that it is the source-side representation engine of the model. We justify this focus in Appendix A where we find that fine-tuning only the encoder achieves similar robustness to fine-tuning the full model, which is not the case when fine-tuning only the decoder or the cross attention.

3.3 Model Analysis

3.3.1 GED Probing

Following (Davis et al., 2022), we perform GED probing to understand how the detection of ungrammatical words changes through the encoder layers. We train linear probes on the word representation of each encoder layer to predict whether the word is grammatically correct or not. Similarly to previous work (Liu et al., 2019; Davis et al., 2022), we take the representation of the last subword as the word representation when a word is split into subwords.

3.3.2 Representation Similarity

To study how encoders affect the representation of the ungrammatical word toward its correct form, we measure the distance between the ungrammatical word and its target grammatical form in each of the encoder layers. We use Centered Kernel Alignment (CKA) (Kornblith et al., 2019) to measure the distance as $1 - CKA(\widetilde{W}, W)$, where $\widetilde{W} \in \mathbb{R}^{N \times d}$ are ungrammatical word representations, $W \in \mathbb{R}^{N \times d}$ are their target grammatical word representations, N is the number of data points, and d is the model hidden dimension. We use CKA because compared to other similarity methods, it does not require the number of data points to be considerably higher than the representation dimension (Kornblith et al., 2019). We note that CKA outputs a similarity score between 0 and 1.

3.3.3 Robustness Heads

3.3.4 From Influential Heads to Robustness Heads

Voita et al. (2019) measured the amount of influence of a token w_i on another token w_j as the distance between w_j's representation before and after w_i was masked. We apply the same method but to measure head influence instead. We collect the word representation after masking one head at a time, then we compute the CKA distance to the original word representation from the same layer. In our hypothesis, masking an attention head that has the most influence on the word's representation will lead to the highest distance, therefore, we term this *Influential Heads*. Formally, for a layer l, we mask each head $h_i \in \{1, 2, ..., H\}$ in layer $l - 1$, and we take the representation of the word w to compute $1 - CKA(w_{h_i}, w)$, where H is the number of heads, w_{h_i} is the word representation after masking the head h_i, and w is the original word representation.

Since we are more interested in understanding which attention heads contribute to correcting the ungrammatical word's representation, we are led to introduce what we term *Robustness Heads*, which we define as heads that influence the ungrammatical word's representation toward its grammatical form. This requires a simple redefinition of *In-*

fluential Heads, where instead of computing the distance to the original word itself, we compute the distance from the noisy word representation to the representation of its clean form. Formally, instead of computing $1-CKA(w_{h_i}, w)$ we compute $1 - CKA(\widetilde{w_{h_i}}, w)$, where $\widetilde{w_{h_i}}$ is the representation of the ungrammatical word $\widetilde{w}$ after head h_i is masked.

3.3.5 Attention to POS Tags

The attention mechanism offers a straightforward way to interpret NLP models, based on the assumption that, like human attention, models focus on parts of the sentence that they find important for making a prediction. This assumption combined with the granularity and clear linguistic functions of the grammatical errors we introduce, offer a way to linguistically analyze the attention of *Robustness Heads*, which we achieve by inspecting their attention to Part Of Speech (POS) Tags. We collect the attention scores directed from the ungrammatical tokens to the other tokens in the sentence, then group the attention scores over words, following (Clark et al., 2019). When the noisy word is split into tokens, we take the mean of the scores. Conversely, when the word it is attending to is split into tokens, we take their sum. This preserves the property that word level attention scores sum to 1. Following this transformation, we use Spacy [2] to label each word in the sentence with its POS tag and collect the attention scores that each POS tag has received.

4 Experimental Setup

4.1 Data

We use the Europarl-ST (Iranzo-Sánchez et al., 2020) dataset which contains official speech, transcriptions and translations of European Parliament debates of multiple European languages. For this work, we only use the transcriptions and translations of 5 directions: En-Es, En-De, En-It, En-Nl and Fr-Es (Dataset splits and sizes are in §B.2). Introducing grammatical errors and interpreting the results of our analysis required understanding of the source languages, which limited the choice of language directions. We were also limited by the performance of models (e.g. MBART scores 51.18 COMET on Fr-Nl) and availability of NLP resources for introducing grammatical errors. Nevertheless, there is a high number of non-native speakers of both English and French, which makes them relevant for our analysis.

After we introduce synthetic errors into the dataset, we sample 30% of the train set for training the GED probes and we use the rest for fine-tuning the models. When sampling, we make sure that the error labels are balanced between the two subsets. Since we seek to experiment with fine-tuning on clean data as well, we keep the clean version of the fine-tuning subset. This results in clean and noisy fine-tuning subsets for the 3 grammatical errors that we introduce, which we use to fine-tune Clean-Finetuned and Noise-Finetuned models of each error type. While for GED probing, we only use the noisy version to probe the models on each error type.

4.2 Synthetic Errors

Nounnum: We find nouns in the sentence then sample one of them and change its number from plural to singular or the opposite depending on its actual number. For English, we use Berkeley parser (Petrov et al., 2006) to identify nouns and their number, and for French we use Spacy-lefff [3].

Article* and *Prep: We find articles or prepositions in the sentence with string matching, and sample one of them uniformly, then sample a replacement based on statistics from CONLL-14 Shared Task on GEC (Ng et al., 2014) dataset in the case of English, and uniformly in the case of French. The list of articles and prepositions is provided in Appendix B.1.

Morpheus: MORPHEUS (Tan et al., 2020) is a black-box adversarial attack that greedily introduces inflectional errors to minimize or maximize a target metric, which in our case is the BLEU score. Similarly to the original work, we only inflect nouns, verbs and adjectives, and we restrict the possible inflections to the original POS tag to preserve the meaning of the sentence. We use Spacy and Spacy-lefff for tokenization and POS tagging of English and French sentences respectively, and we use LemmInflect [4] to find inflections for English similarly to the original work, and Inflecteur [5] to find inflections for French. For further details, we refer the reader to the original work (Tan et al., 2020).

[2] https://spacy.io/

[3] https://github.com/sammous/spacy-lefff

[4] https://github.com/bjascob/LemmInflect

[5] https://github.com/Achuttarsing/inflecteur

4.3 Models

We run experiments on multilingual machine translation models. More specifically, we analyzed three models: Multilingual translation version of MBART (Tang et al., 2020), M2M100 (Fan et al., 2021) and NLLB (NLLB Team et al., 2022). We also experimented with bilingual models of OPUS-MT (Tiedemann and Thottingal, 2020) for En-Es, En-De, En-It, En-Nl and Fr-Es. Each of these models has a different level of multilinguality, where OPUS-MT models support a single direction, while MBART, M2M100 and NLLB support many-to-many translation between 50 languages, 100 languages, and over 200 languages respectively.

4.4 Fine-tuning

We fine-tune the models using HuggingFace transformers library [6]. We use AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 5e-05 and a batch size of 64. We save the best model based on development set BLEU score during a maximum of 5000 steps. For validation, we use the clean and noisy development sets when fine-tuning on the clean and noisy subsets respectively. We fine-tune the models with a frozen decoder unless it is mentioned otherwise.

4.5 GED Probe Training

We train single layer probes using the Pytorch framework. we use Adam optimizer with a learning rate of 1e-03 and a weight decay of 1e-04. We train with a batch size of 32 for 50 epochs with a dropout of 0.1 and a patience of 10 epochs for early stopping based on validation F1 score.

4.6 Evaluation

We evaluate translation on the Europarl-ST test sets. For each grammatical error, we report results on both clean and noisy test sets. We also report the difference in performance between clean and noisy. As metrics, we use COMET (Rei et al., 2020), BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). COMET is a neural based metric that was shown to be more aligned with human judgments (Freitag et al., 2022). We use the reference-based model wmt22-comet-da [7], and we report BLEU and ChrF results in our repository.

For GED probing, we evaluate probes of each model on the noisy version of the test set of each grammatical error using the F1 score.

[6] https://github.com/huggingface/transformers

[7] https://huggingface.co/Unbabel/wmt22-comet-da

5 Results and Analysis

5.1 Fine-tuning

In Table 1 we show the COMET scores of Base, Clean-Finetuned and Noise-Finetuned models on clean and noisy test sets and their difference (Δ) for En-Es (and Table 3 for the other language directions). We can see that even in our simple setup, where we insert one error per sentence, we still see a significant drop in performance in Base models (0.66 at minimum) (Kocmi et al., 2024). Furthermore, fine-tuning on clean or noisy data leads to better results but only fine-tuning on noisy data leads to improving robustness, which is seen in the reduced difference in COMET (e.g. from 0.74 to 0.01 for NLLB on *Article* errors). Surprisingly, fine-tuning on grammatical errors doesn't affect performance on clean data, and can even lead to better results compared to fine-tuning on clean data (e.g. 77.51 compared to 77.14 for M2M100 on *Prep* errors). This suggests that fine-tuning on grammatical errors has a regularization effect. Across error types, we see that *Prep* errors lead to the most significant drop, and that even after fine-tuning, the drop in performance is still significant (e.g. up to 0.29 for NLLB). We note that Clean-Finetuned models performance on clean data is different across errors because the clean train subsets were sampled to match the noisy subsets which are different because of the error distribution.

In this sub-section, we established that grammatical errors lead to a significant drop in performance, and that fine-tuning on them leads to increased robustness, therefore in the next sub-section we proceed to analyze Base, Noise-Finetuned and Clean-Finetuned models representations when responding to the grammatical errors. Furthermore, in Appendix C.6 we provide the results on *Morpheus*, which confirms the generalizability of our analysis to other error types and to sentences containing more than one error per sentence.

5.2 GED Probing

Figure 1 (and Figure 6 for the other language directions) shows that the GED probing performance improves during roughly the first half layers of the model, then generally plateaus in the second half for Base and Clean-Finetuned but decreases for Noise-Finetuned models. Additionally, across Base models the representation of errors achieves

Direction	Model	Article			Nounnum			Prep		
		Clean	Noisy	Δ	Clean	Noisy	Δ	Clean	Noisy	Δ
En-Es	opus-mt-base	78.72	77.71	1.0	78.72	77.74	0.97	78.72	77.67	1.05
	opus-mt-clean	78.88	78.05	0.84	78.97	78.12	0.85	78.93	78.05	0.88
	opus-mt-noise	**78.94**	**78.89**	**0.06**	**78.99**	**78.82**	**0.17**	**79.06**	**78.83**	**0.23**
	m2m100-base	75.99	74.84	1.15	75.99	75.12	0.88	75.99	74.63	1.36
	m2m100-clean	77.39	76.4	0.99	77.28	76.22	1.07	77.14	76.14	1.0
	m2m100-noise	**77.57**	**77.54**	**0.03**	**77.58**	**77.5**	**0.08**	**77.51**	**77.23**	**0.28**
	mbart-base	78.04	77.23	0.81	78.04	77.25	0.79	78.04	77.24	0.79
	mbart-clean	78.51	77.77	0.74	**78.64**	77.74	0.9	78.49	77.73	0.76
	mbart-noise	**78.61**	**78.58**	**0.04**	78.57	**78.51**	**0.06**	**78.58**	**78.38**	**0.2**
	nllb-base	78.34	77.6	0.74	78.34	77.69	0.66	78.34	77.52	0.83
	nllb-clean	**78.82**	78.14	0.68	**78.85**	78.16	0.69	78.81	78.21	0.6
	nllb-noise	78.81	**78.8**	**0.01**	78.78	**78.73**	**0.06**	**78.94**	**78.66**	**0.29**

Table 1: COMET scores on En-Es. The Base model is the original model, Clean is fine-tuned on the clean version of the data, and Noise is fine-tuned on the noisy version (with the same noise as the one they are tested on). We present the performance on the clean and noisy test sets and their difference (Δ).

closely similar GED F1 score on each layer, while across errors, the F1 scores are different. For example, the GED F1 score of *Nounnum* errors at layer 1 is almost 0., while that of *Article* errors is 0.7. When looking at French as a different source language, Figure 6(a) shows that French error detection is represented differently, especially that of *Nounnum* errors where the maximum F1 score on En-Es is 0.48, while on Fr-Es it is 0.84. This might be explained by the fact that in French, the noun number is indicated by adjectives and articles, while it is not the case for English.

Fine-tuning negatively affects the GED probing performance in deeper layers, where our GEC setup hypothesis suggests that correction is happening. Noise-Finetuned models maintain their ability to detect errors in lower layers, then correct them in deeper layers, which makes the GED probing accuracy more challenging because the representation is corrected. In the next sub-section, we further support this correction hypothesis.

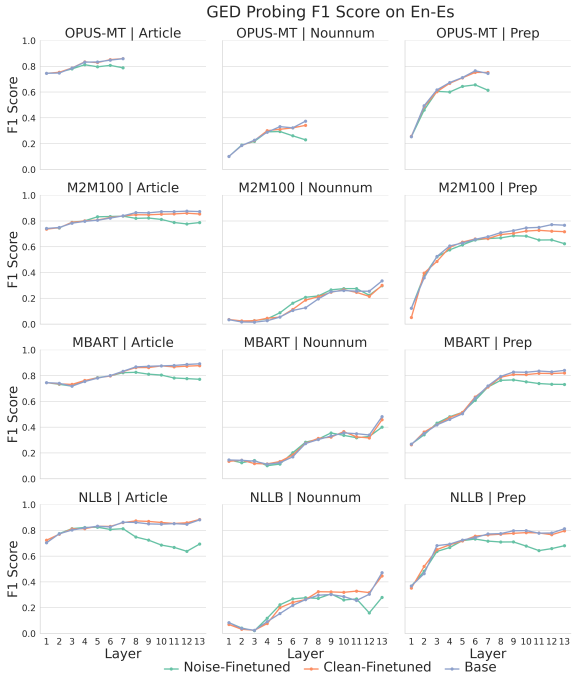


Figure 1: GED probing performance of Noise-Finetuned, Clean-Finetuned and Base models on En-Es. GED probing performance of Noise-Finetuned models witnesses a degradation in deeper layers.

5.3 Representation Distance

Figure 2 (and Figure 7 for other language directions) shows the representation distance between the ungrammatical or noisy word and its target clean word at each layer of the encoder. Generally, for Base and Clean-Finetuned models the CKA distance decreases from one layer to the next except for *Prep* errors where the distance decreases then increases in deeper layers. This can be explained by the fact that both words have the same linguistic function (nouns, articles or prepositions), and because they share the same context which leads their representation to move closer as the encoder integrates context into it. On the other hand, Noise-Finetuned models exhibit similar behavior to their Base model but they learn to drive the representation to be closer (almost 0. CKA distance in most cases), this means that the similarity is driven by robustness as well, where models correct ungrammatical words by pushing their representation toward their grammatical form. Combined with the GED probing results, this supports our hypothesis that the encoder detects and corrects grammatical errors to achieve robustness. In the next sub-section, we analyze *Robustness Heads* to explain this behavior in Noise-Finetuned Models.

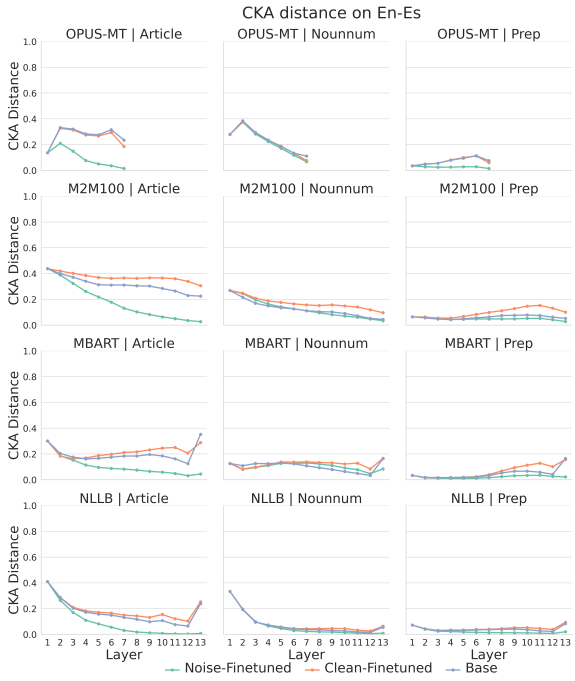


Figure 2: CKA distance of clean and noise word representations across models and errors on En-Es. Noise-Finetuned models drive the representation of the noisy word to be more similar to the clean word.

5.4 Robustness Heads

5.4.1 Attention to POS Tags

Figure 3 (and Figures 8, 9, 10, 11 for other language directions) presents the average attention scores of each POS tag. The scale of attention is relative to each base model and error type. For clarity reasons, we only show the attention scores over the 10 most common POS tags in the dataset. We keep the tags as they are named in Spacy, but their full names are presented in Appendix B.3. The figure shows that generally, the attention of *Robustness Heads* to POS tags is concentrated in the early layers which is related to how early work have found that lower layers are better at POS tagging (Belinkov et al., 2017). Furthermore, *Robustness Heads* attend to words in the sentence that can help identify or correct the grammatical error in question. Although the attention is distributed differently across models, they still attend to similar POS tags when responding to the same error. However, when comparing between English and French, models attend to different POS tags, which can be explained by their linguistic differences. If we look at *Article* errors, English models primarily focus their attention on adjectives, nouns, and proper nouns, while their French counterparts primarily focus on nouns. This difference can be attributed to the ordering of adjectives in each language. In English, adjectives mostly precede nouns, while the opposite is true for French. This means that adjectives in English have more direct influence on articles, especially when making the choice between "a" and "an" (they depend on whether the next word starts with a consonant or a vowel). Another fact that could lead the model to focus less on adjectives in French, is that certain adjectives do not follow their noun in gender or number. When examining *Nounnum* errors, models exhibit a common pattern of focusing primarily on adjectives and determiners, however, attention to determiners in English is notably lower compared to French, which can be explained by the fact that in French, the noun number is indicated by articles. Finally, for *Prep* errors, the models focus mainly on verbs, nouns, and determiners. The three definitely can affect the choice of prepositions; although it is not clear with determiners, we note that some prepositions are more common with definite vs. indefinite determiners, such as "on" and "the".

Base, Clean-Finetuned, and Noise-Finetuned models distribute their attention similarly to POS tags, but in some cases, the noisy model has learned to put more attention to the correct POS tags. For example, M2M100 model on Fr-Es has learned to put more attention on determiners and adjectives to deal with *Nounnum* errors (going from a mean over layers of 0.021 and 0.030 to 0.023 and 0.035 respectively). On Fr-Es *Article* errors, NLLB and OPUS-MT models learned to put more attention on nouns (from 0.051 and 0.065 to 0.053 and 0.069). On En-Es *Article* errors, NLLB and OPUS-MT have learned to put more attention on nouns and proper nouns respectively (going from 0.052 and 0.057 to 0.055 and 0.061), while on *Prep* errors, they have learned to put more attention on verbs (from 0.062 and 0.062 to 0.067 and 0.069). We see this trend in other language directions as well as across models and errors, although in some cases it is not very clear and the attention scores of *Robustness Heads* of Noise-Finetuned and Base models are very close.

5.4.2 Similarity between Robustness and Influential Heads

The previous section shows that Base, Clean-Finetuned and Noise-Finetuned models each contain *Robustness Heads* that attend similarly to interpretable POS tags when dealing with grammatical errors. So what gives the Noise-Finetuned

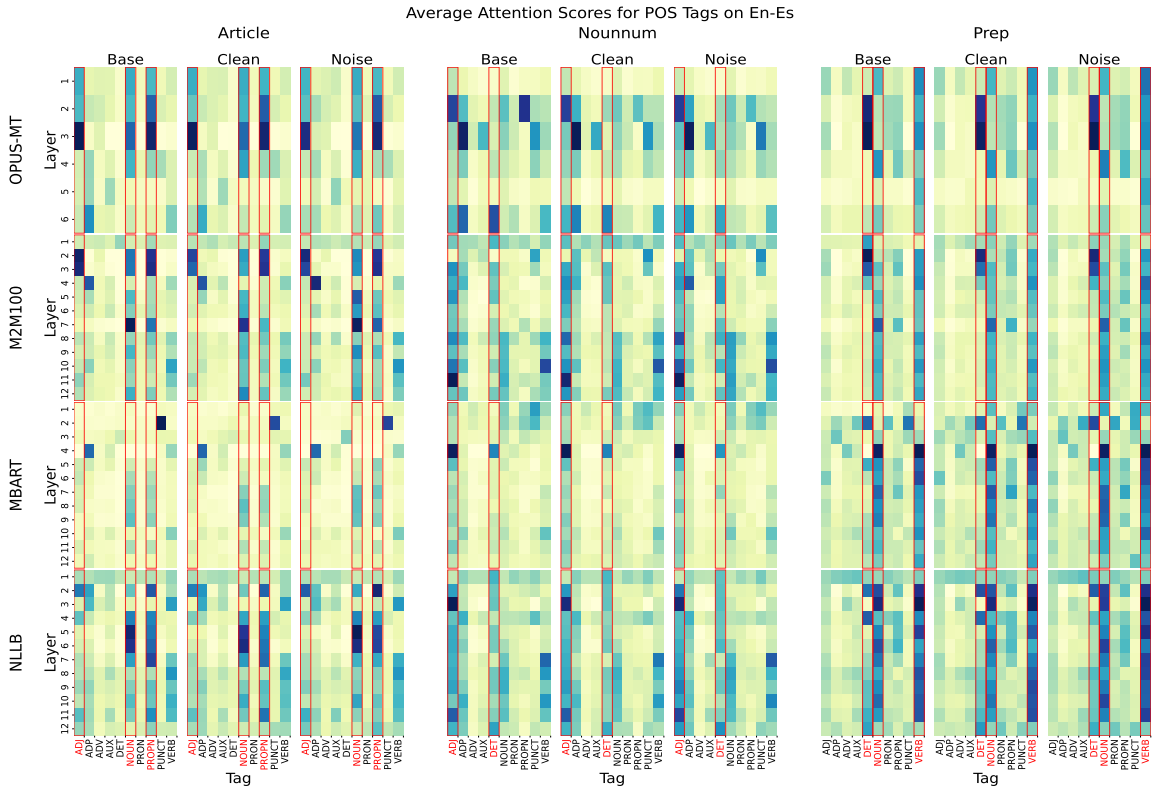


Figure 3: *Robustness Heads* attention to the 10 most common POS tags in the test set on En-Es. The scale of attention is relative to each base model and error. We highlight POS tags that are attended to the most across models.

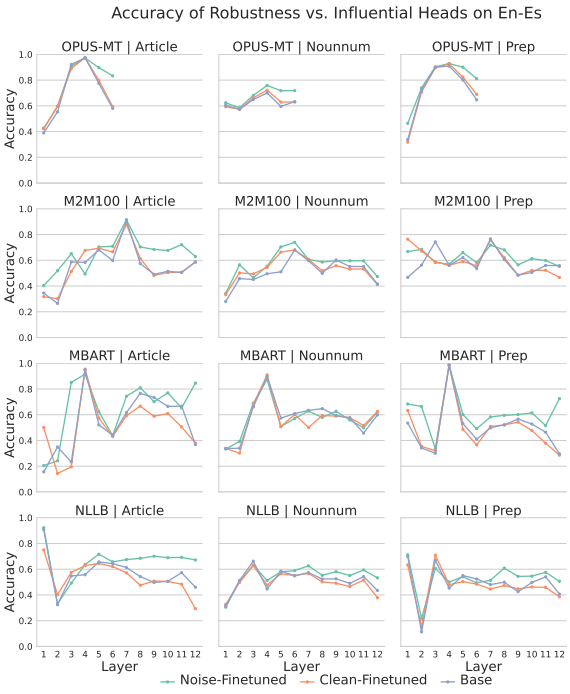


Figure 4: Accuracy of Robustness and Influential heads on En-Es. We find the accuracy is higher for Noise-Finetuned models especially in deep layers.

models an advantage in terms of robustness? To answer this question, we start by identifying *Influential Heads* of the noisy word, as the heads that most influence it toward its current state, then we compare them with *Robustness Heads*. Figure 4 presents the accuracy between *Influential Heads* and *Robustness Heads* at each layer of the encoder. The figure shows that this accuracy is higher in Noise-Finetuned models especially in deeper layers, which means that models after fine-tuning on noise, tend to employ more *Robustness Heads* for updating the noisy word representation, and this can explain their improved robustness.

6 Discussion

RQ1: *How do models represent and handle grammatical errors to achieve robustness?* The encoder implements a GEC setup where it first detects the error then corrects it, and this behavior is more distinguishable in Noise-Finetuned models. Compared to Base or Clean-Finetuned models, Noise-Finetuned models maintain their error detection in lower layers, while they drive the representation of the ungrammatical word to be as closely similar to the grammatical form (§5.3). We argue that the correction component becomes more prominent later in the model, leading to lower GED probing accuracy and increased usage of *Robustness Heads*

in deeper layers (§5.2 and §5.4.2). This finding coupled with the results in Appendix A show that fine-tuning only the encoder while freezing the decoder is sufficient for achieving robustness, while also preserving performance on clean data.

RQ2: ***How does grammatical error representation differ across models and languages?*** Davis et al. (2022) suggest that GED probing performance can reflect the linguistic knowledge of pre-trained models. While this might be true, our evidence indicates that it is more influenced by the linguistic features of the language itself (specifically the source language) (§5.2). For example, the GED probing performance of *Nounnum* errors on Fr-Es peaks at around 0.8, while on En-Es, it peaks at maximum 0.5. This difference can be explained by the fact that in French articles and adjectives follow their noun in number, providing easy access to information about the noun in question. Furthermore, we find similarities across models in our analysis of representation distance and attention to POS tags as well (§5.3 and §5.4.1). However, when comparing across source languages, models represent and handle the same error differently. This suggests that languages might interfere with one another when fine-tuning for robustness across multiple languages, or even during training, which can be avoided by using language-specific adapters (Bapna and Firat, 2019).

RQ3: ***How does fine-tuning for robustness influences models to achieve improved robustness?*** Fine-tuning for robustness leverages the existing knowledge of the Base models and generally does not go beyond it. If we look at probing performance, we see that generally, Noise-Finetuned models do not go beyond the peak in performance achieved by their Base model (§5.2), even though they are trained on the grammatical errors. Moreover, the *Robustness Heads* of Noise-Finetuned models distribute their attention to POS tags similarly to their Base models (§5.4.1). However, we find that Noise-Finetuned models rely more on *Robustness Heads* especially in deeper layers to influence the noisy word toward the correct form (§5.4.2). This suggests that fine-tuning for robustness builds on the existing structure in pre-trained NMT models, therefore, analyzing the effects of different pre-training strategies and training data can be a valuable direction for future work.

7 Conclusion and Future Work

In this work, we analyze transformer NMT encoders under the effect of grammatical noise, and investigate how fine-tuning for robustness affects model behavior and internal representations. We find that the encoder -especially after fine-tuning- implements a GEC setup: it first detects the error and then corrects it by adjusting the representation of the ungrammatical word towards its correct form. To better understand this behavior, we propose a method for finding *Robustness Heads*- attention heads that attend to POS tags and help detect and correct the grammatical error. Additionally, we find that fine-tuning on grammatical errors leads the model to use more *Robustness Heads* especially in deeper layers.

These findings suggest a practical strategy for improving robustness in NMT systems: fine-tuning only the encoder on (synthetically) noisy data can substantially enhance robustness without degrading performance on clean data. This makes it an efficient and interpretable alternative to full model fine-tuning. In practice, such systems may benefit from selectively introducing common error types—especially those relevant to their deployment context—during fine-tuning.

Furthermore, our analysis framework, which combines GED probing, representational similarity and influential attention heads, is model-agnostic and generalizable. It provides a systematic way to audit robustness across different models and languages and could be applied to other encoder-decoder NLP systems.

Although we focused our analysis on encoder-decoder models, we hypothesize that decoder-only models may exhibit a similar behavior especially in their early layers, therefore, we will extend this analysis to decoder only models in future work.

8 Limitations

Although Spanish, Italian and French belong to the Romance family, while English, German, and Dutch are Germanic languages. All of the languages belong to the Indo-European family, which might limit the generalization of our work to other languages. Additionally, we use synthetic noise which is more controllable and allows us to provide a fine-grained analysis of models, but there are definitely limitations in how close it can simulate natural noise. Finally, due to resource limitations, we focus on relatively small models, and while

larger models might be more robust, we think that our analysis still offers interesting insights about robustness.

Acknowledgments

The research presented in this paper was conducted as part of VOXReality project[8], which was funded by the European Union Horizon Europe program under grant agreement No 101070521.

References

Antonios Anastasopoulos. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 213–223, Florence, Italy. Association for Computational Linguistics.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Junkai Chen, Zhenhao Li, Xing Hu, and Xin Xia. 2024. Nlperturbator: Studying the robustness of code llms to natural language variations. *Preprint*, arXiv:2406.19783.

[8] https://voxreality.eu/

Christian Cianfarani, Arjun Nitin Bhagoji, Vikash Sehwag, Ben Zhao, Haitao Zheng, and Prateek Mittal. 2022. Understanding robust learning through the lens of representation similarities. In *Advances in Neural Information Processing Systems*.

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.

Christopher Davis, Christopher Bryant, Andrew Caines, Marek Rei, and Paula Buttery. 2022. Probing for targeted syntactic knowledge through grammatical error detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 360–373, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2023a. Discovering salient neurons in deep nlp models. *Journal of Machine Learning Research*, 24(362):1–40.

Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2023b. Discovering salient neurons in deep nlp models. *Journal of Machine Learning Research*, 24(362):1–40.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020.

Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The overview of the SST speech corpus of Japanese learner English and evaluation through the experiment on automatic detection of learners' errors. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sai Muralidhar Jayanthi and Adithya Pratapa. 2021. A study of morphological robustness of neural machine translation. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–59, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. *CoRR*, abs/1905.00414.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2900–2907, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Courtney Napoles, Aoife Cahill, and Nitin Madnani. 2016. The effect of multiple grammatical errors on processing non-native writing. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, San Diego, CA. Association for Computational Linguistics.

Pavan Kalyan Reddy Neerudu, Subba Oota, Mounika Marreddy, Venkateswara Kagita, and Manish Gupta. 2023. On robustness of finetuned transformer-based NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7180–7195, Singapore. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht,

Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Peyman Passban, Puneeth Saladi, and Qun Liu. 2021. Revisiting robust neural machine translation: A transformer case study. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3831–3840, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Jason Phang, Haokun Liu, and Samuel R. Bowman. 2021. Fine-tuned transformers show clusters of similar representations across layers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 529–538, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

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.

Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech 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 2786–2792, Minneapolis, Minnesota. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Raúl Vázquez, Hande Celikkanat, Mathias Creutz, and Jörg Tiedemann. 2021. On the differences between BERT and MT encoder spaces and how to address them in translation tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 337–347, Online. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Tao Wang, Chengqi Zhao, Mingxuan Wang, Lei Li, Hang Li, and Deyi Xiong. 2021. Secoco: Self-correcting encoding for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4639–4644, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yibin Wang, Yichen Yang, Di He, and Kun He. 2023. Robustness-aware word embedding improves certified robustness to adversarial word substitutions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 673–687, Toronto, Canada. Association for Computational Linguistics.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity analysis of contextual word representation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655, Online. Association for Computational Linguistics.

Weiwen Xu, Ai Ti Aw, Yang Ding, Kui Wu, and Shafiq Joty. 2021. Addressing the vulnerability of NMT in input perturbations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 80–88, Online. Association for Computational Linguistics.

Yichen Yang, Xiaosen Wang, and Kun He. 2022. Robust textual embedding against word-level adversarial attacks. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Jingyi Zhang, Gerard de Melo, Hongfei Xu, and Kehai Chen. 2023. A closer look at transformer attention for multilingual translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 496–506, Singapore. Association for Computational Linguistics.

Yuting Zhao and Ioan Calapodescu. 2022. Multimodal robustness for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8505–8516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 565–571, Florence, Italy. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Preprint*, arXiv:2306.04528.

A Freezing Results

When fine-tuning different parts of the model on grammatical errors, while keeping the other parts frozen, we find that fine-tuning only the encoder approaches fine-tuning the full model (both encoder and decoder) in terms of robustness, as shown in Figure 5. Although the target translation is generated on the decoder side, fine-tuning the decoder does not lead to similar improvements as fine-tuning the encoder. This suggests that in NMT, robustness is more related to the source side representation than to the generation process, and therefore, we focus our analysis on the encoder.

B Experiments Details

B.1 Articles and Prepositions Lists

We provide the list of articles and prepositions that we consider for our analysis:
English Articles: {a, an, the}.
English Prepositions: {on, in, at, from, for, under, over, with, into, during, until, against, among, throughout, of, to, by, about, like, before, after, since, across, behind, but, out, up, down, off}.
French Articles: {la, le, un, une, les, des}.
French Prepositions: {à, après, avant, avec, chez, contre, dans, de, depuis, derrière, devant, durant, en, entre, envers, environ, jusque, malgré, par, parmi, pendant, pour, sans, sauf, selon, sous, suivant, sur, vers}.

B.2 Data Splits

Table 2 shows the amount of data for each language direction in our experiments. Although the train set for Fr-Es is small in comparison to the other language directions, the COMET improvements after fine-tuning shown in Table C.1 suggest that it is enough for fine-tuning and potentially for drawing conclusions about its effects. The test sets which we use for evaluating models and for our analysis are almost similar in size.

B.3 POS Tags List

ADJ: Adjective.
ADP: Adposition.
ADV: Adverb.

	Train	Dev	Test
En-Es	31607	1272	1267
En-De	32628	1320	1253
En-Nl	31401	1269	1235
En-It	29552	1122	1130
Fr-Es	7857	1072	1098

Table 2: Dataset splits for each language direction

AUX: Auxiliary.
CCONJ: Coordinating conjunction.
DET: Determiner.
NOUN: Noun.
PRON: Pronoun.
PROPN: Proper noun.
PUNCT: Punctuation.
VERB: Verb.

C Additional Results

C.1 Fine-tuning Results

We present the results of fine-tuning NMT models for robustness on Fr-Es, En-De, En-It and En-Nl in Table 3.

C.2 GED Probing

Figure 6 shows the GED probing performance on Fr-Es, En-De, En-It, and En-Nl.

C.3 Respresentation Similarity

Figures 7 shows the CKA distance of clean and noisy word representations on Fr-Es, En-De, En-It and En-Nl.

C.4 Attention to POS Tags

Figure shows the accuracy of Robustness and Influential heads on *Morpheus* errors. Figures 8, 9, 10 and 11 show *Robustness Heads* attention to POS tags on Fr-Es, En-De, En-It and En-Nl respectively.

C.5 Similarity between Robustness and Influential Heads

Figure 12 shows the accuracy of Robustness and Influential heads on Fr-Es, En-De, En-It and En-Nl.

C.6 Generalization to Morpheus

To validate the generalization of our analysis, we present the results of using Morpheus. We find similar results in terms of fine-tuning performance, probing, representational distance and similarity between robustness and influential heads (Shown in Table 4 and Figures 13, 14 and 15 respectively). We do not present the results of attention to POS tags because the interpretation of results requires granular errors.

C.6.1 Fine-tuning Results

We present the results of fine-tuning for robustness to *Morpheus* errors in Table C.6.

C.6.2 GED Probing

Figure 13 shows the GED probing performance of *Morpheus* errors.

C.6.3 Representation Distance

Figures 14 shows the CKA distance of clean and noisy word representations of *Morpheus* errors.

C.6.4 Similarity between Robustness and Influential Heads

Figure 15 shows the accuracy of Robustness and Influential heads on *Morpheus* errors.

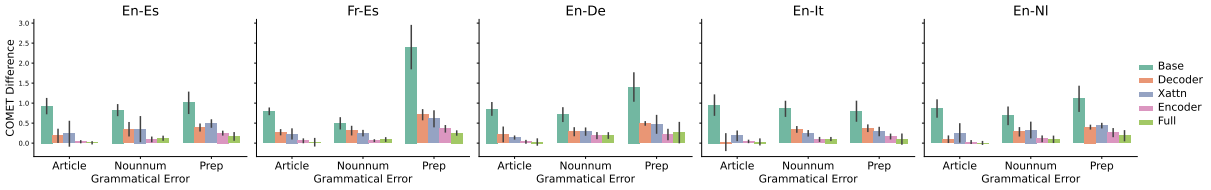
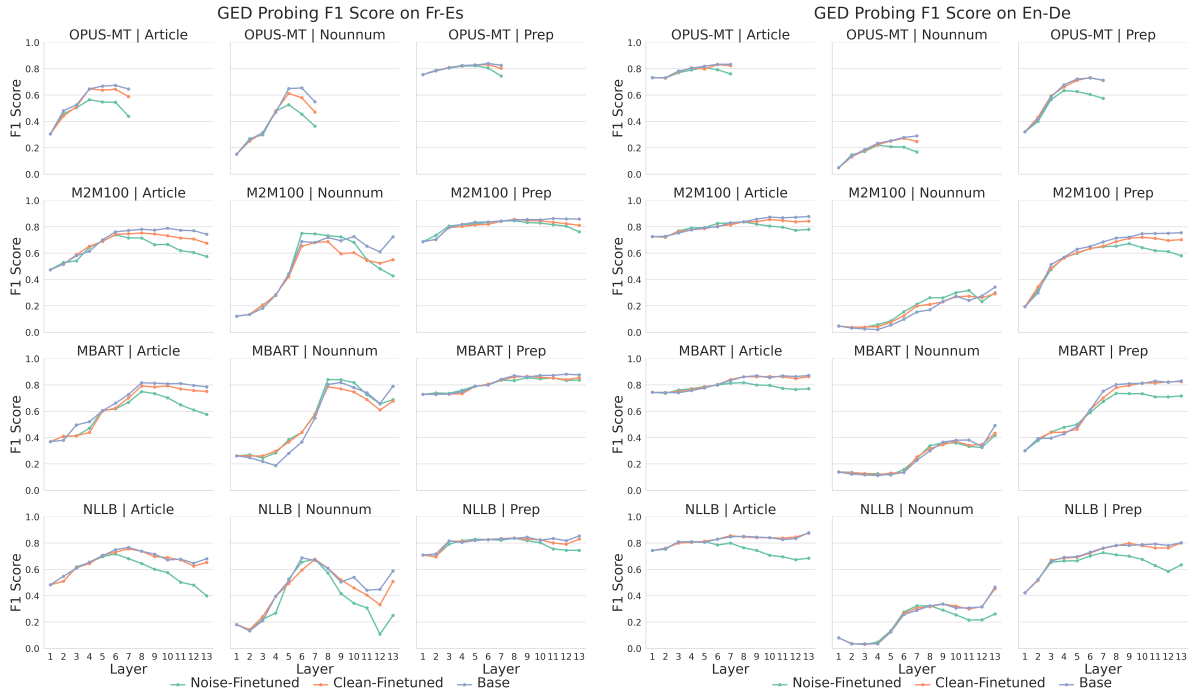


Figure 5: COMET difference between performance on clean and noisy test sets after fine-tuning different parts of the models. We show the average across 3 multilingual models (M2M100, MBART, NLLB) and bilingual models (OPUS-MT for each language pair). Fine-tuning the encoder is almost as good as fine-tuning the full model.

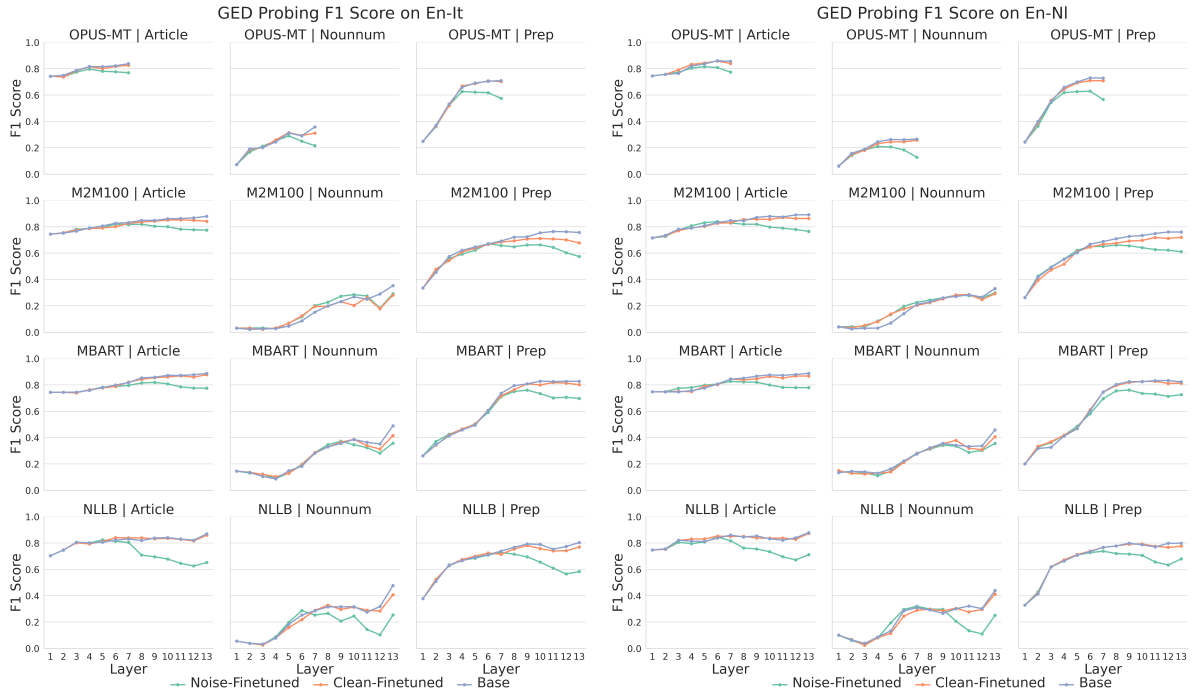
		Article			Nounnum			Prep		
Direction	**Model**	**Clean**	**Noisy**	**Δ**	**Clean**	**Noisy**	**Δ**	**Clean**	**Noisy**	**Δ**
Fr-Es	opus-mt-base	74.57	73.79	0.79	74.57	73.98	0.6	74.57	71.88	2.7
	opus-mt-clean	75.36	74.7	0.66	75.13	74.7	0.43	**75.7**	73.11	2.58
	opus-mt-noise	**75.62**	**75.54**	**0.07**	**75.69**	**75.62**	**0.07**	75.65	**75.18**	**0.46**
	m2m100-base	73.04	72.21	0.83	73.04	72.45	0.59	73.04	70.06	2.98
	m2m100-clean	74.01	73.45	0.55	73.98	73.65	0.33	**74.27**	72.21	2.07
	m2m100-noise	**74.24**	**74.23**	**0.01**	**74.22**	**74.16**	**0.06**	74.2	**73.84**	**0.36**
	mbart-base	69.13	68.27	0.87	69.13	68.82	0.31	69.13	67.02	2.12
	mbart-clean	**74.25**	**73.72**	0.52	74.28	73.7	0.58	74.17	72.41	1.76
	mbart-noise	73.68	73.57	**0.11**	**74.33**	**74.29**	**0.04**	**74.23**	**73.9**	**0.33**
	nllb-base	74.92	74.23	0.69	74.92	74.44	0.48	74.92	73.12	1.8
	nllb-clean	**76.1**	75.55	0.55	76.03	75.62	0.41	75.99	74.68	1.31
	nllb-noise	76.0	**75.97**	**0.04**	**76.12**	**76.04**	**0.08**	**76.07**	**75.79**	**0.28**
En-De	opus-mt-base	76.93	75.89	1.04	76.93	76.0	0.93	76.93	75.22	1.71
	opus-mt-clean	77.83	76.91	0.93	77.82	77.03	0.79	77.89	76.47	1.42
	opus-mt-noise	**77.86**	**77.86**	**0.0**	**77.93**	**77.73**	**0.2**	**77.93**	**77.54**	**0.4**
	m2m100-base	72.94	72.05	0.89	72.94	72.19	0.75	72.94	71.42	1.52
	m2m100-clean	75.29	74.64	0.65	75.08	74.31	0.78	75.1	74.02	1.09
	m2m100-noise	**75.61**	**75.54**	**0.07**	**75.42**	**75.14**	**0.28**	**75.31**	**75.14**	**0.17**
	mbart-base	76.26	75.44	0.82	76.26	75.64	0.62	76.26	74.78	1.48
	mbart-clean	77.29	76.84	0.45	77.35	76.87	0.48	**77.54**	76.66	0.88
	mbart-noise	**77.5**	**77.49**	**0.02**	**77.5**	**77.35**	**0.15**	77.47	**77.34**	**0.13**
	nllb-base	76.7	76.03	0.67	76.7	76.15	0.55	76.7	75.79	0.9
	nllb-clean	77.52	76.96	0.56	77.49	77.05	0.43	77.6	76.72	0.88
	nllb-noise	**77.56**	**77.52**	**0.04**	**77.51**	**77.39**	**0.12**	**77.61**	**77.45**	**0.16**
En-It	opus-mt-base	77.22	76.08	1.14	77.22	76.26	0.95	77.22	76.13	1.09
	opus-mt-clean	77.31	76.35	0.96	77.45	76.53	0.92	77.42	76.37	1.05
	opus-mt-noise	**77.53**	**77.46**	**0.07**	**77.71**	**77.58**	**0.13**	**77.7**	**77.51**	**0.2**
	m2m100-base	75.0	73.82	1.19	75.0	73.95	1.06	75.0	74.11	0.89
	m2m100-clean	**76.83**	75.77	1.06	76.5	75.51	0.99	76.61	75.88	0.73
	m2m100-noise	76.81	**76.8**	**0.01**	**76.83**	**76.8**	**0.03**	**76.91**	**76.75**	**0.16**
	mbart-base	75.73	74.97	0.75	75.73	75.08	0.65	75.73	75.04	0.68
	mbart-clean	**77.63**	76.88	0.76	**77.75**	77.01	0.74	**77.67**	77.18	0.48
	mbart-noise	77.58	**77.53**	**0.05**	77.73	**77.62**	**0.11**	77.62	**77.41**	**0.21**
	nllb-base	77.47	76.75	0.72	77.47	76.7	0.77	77.47	76.93	0.53
	nllb-clean	**78.07**	77.45	0.61	**77.94**	77.35	0.59	**77.94**	77.53	0.41
	nllb-noise	77.96	**77.92**	**0.04**	77.88	**77.77**	**0.1**	77.86	**77.76**	**0.09**
En-Nl	opus-mt-base	78.03	76.99	1.03	78.03	77.13	0.89	78.03	76.62	1.41
	opus-mt-clean	78.66	77.76	0.9	78.65	77.83	0.82	78.86	77.61	1.25
	opus-mt-noise	**79.05**	**79.0**	**0.06**	**78.93**	**78.8**	**0.14**	**78.98**	**78.62**	**0.36**
	m2m100-base	74.9	73.91	0.98	74.9	74.08	0.82	74.9	73.59	1.31
	m2m100-clean	76.67	75.82	0.86	76.75	75.89	0.86	76.85	75.93	0.92
	m2m100-noise	**76.78**	**76.79**	**-0.01**	**76.9**	**76.81**	**0.08**	**76.99**	**76.66**	**0.32**
	mbart-base	74.86	73.97	0.89	74.86	74.26	0.6	74.86	73.89	0.97
	mbart-clean	**77.82**	77.12	0.71	77.86	77.08	0.79	78.05	77.22	0.83
	mbart-noise	77.77	**77.73**	**0.04**	**77.98**	**77.8**	**0.18**	**78.08**	**77.93**	**0.15**
	nllb-base	77.95	77.4	0.56	77.95	77.53	0.42	77.95	77.21	0.74
	nllb-clean	**78.85**	78.26	0.59	**78.81**	78.3	0.45	**78.74**	78.12	0.63
	nllb-noise	78.74	**78.72**	**0.02**	78.7	**78.67**	**0.03**	78.73	**78.52**	**0.22**

Table 3: COMET scores on Fr-Es, En-De, En-It and En-NL. The Base model is the original model, Clean is fine-tuned on the clean version of the data, and Noise is fine-tuned on the noisy version (with the same noise as the one they are tested on). We present the performance on the clean and noisy test sets and their difference (Δ).



(a) Fr-Es

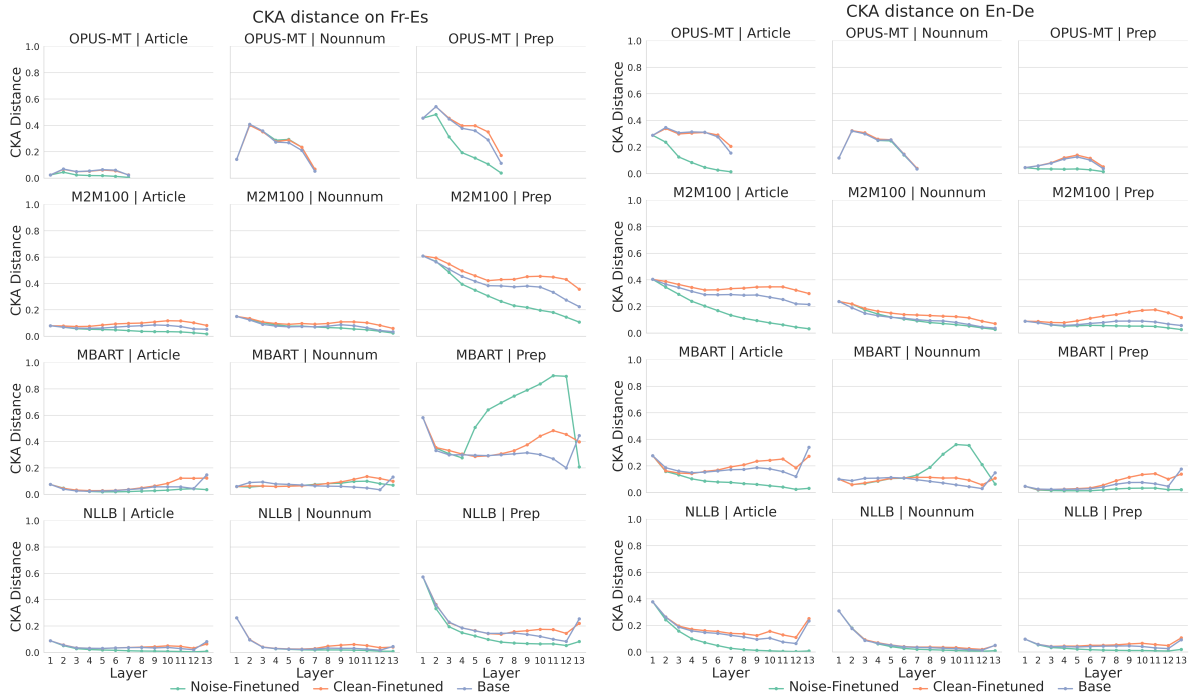
(b) En-De



(c) En-It

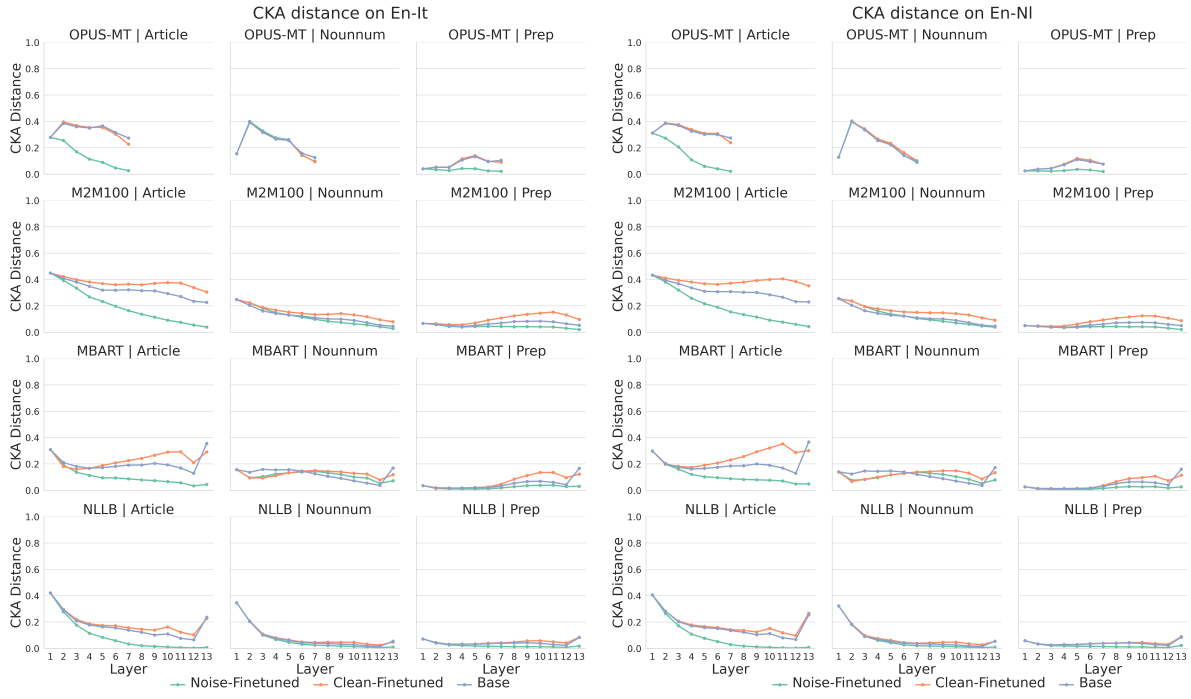
(d) En-Nl

Figure 6: GED probing performance of Noise-Finetuned, Clean-Finetuned and Base models on Fr-Es, En-De, En-It and En-Nl. GED probing performance of Noise-Finetuned models witnesses a degradation in deeper layers.



(a) Fr-Es

(b) En-De



(c) En-It

(d) En-Nl

Figure 7: CKA distance of clean and noise word representations across models and errors on Fr-Es, En-De, En-It and En-Nl. Noise-Finetuned models drive the representations of the noisy word to be more similar to the clean word.

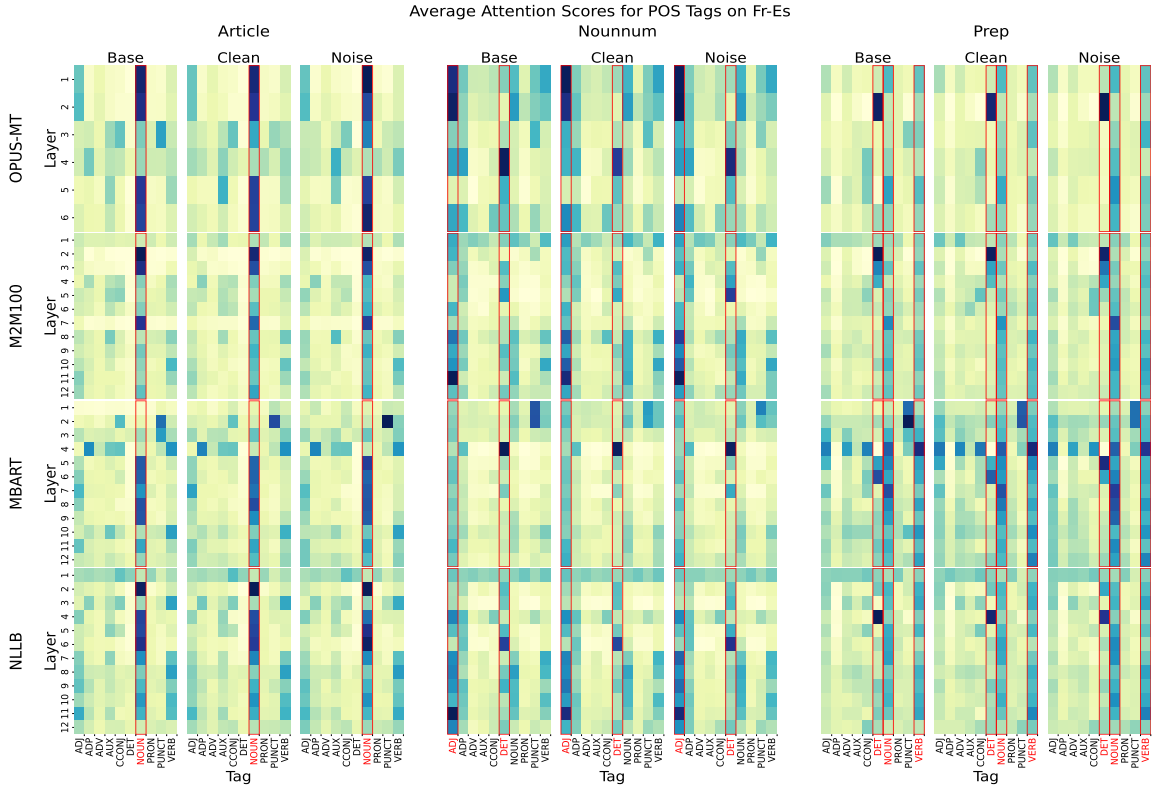


Figure 8: *Robustness Heads* attention to the 10 most common POS tags in the test set on Fr-Es. The scale of attention is relative to each base model and error. We highlight POS tags that are attended to the most across models.

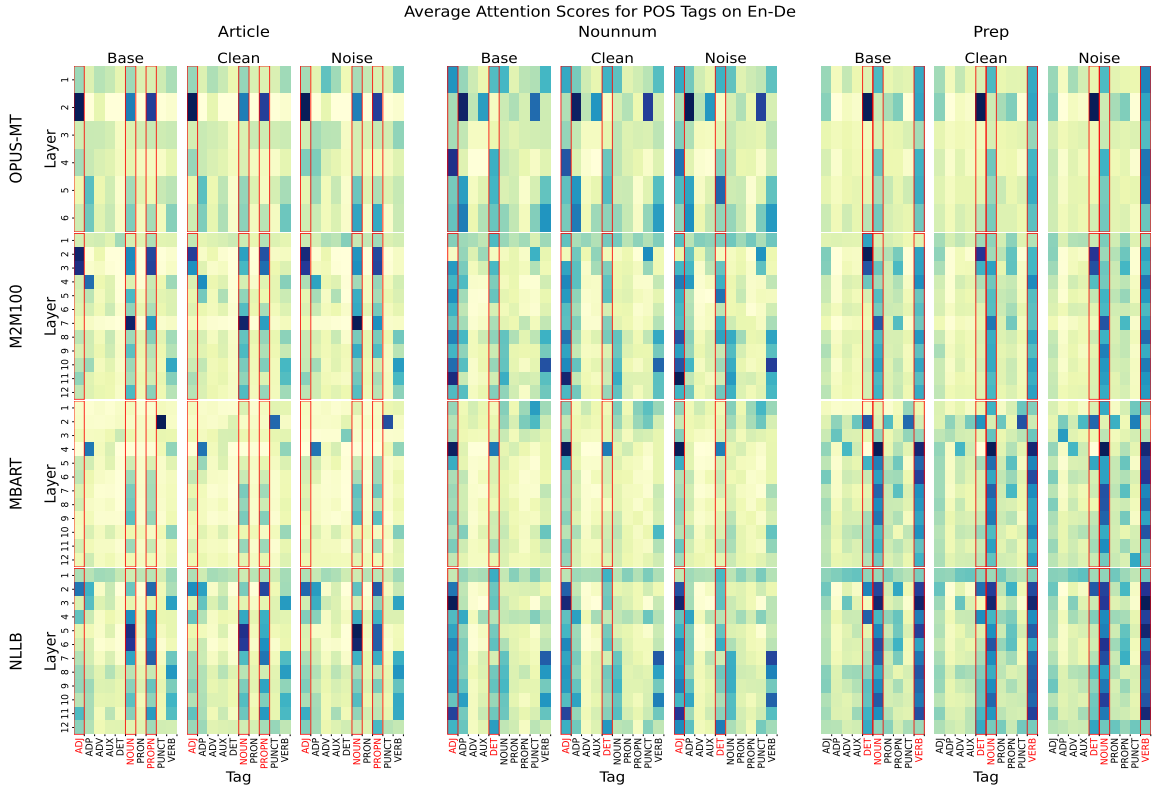


Figure 9: *Robustness Heads* attention to the 10 most common POS tags in the test set on En-De. The scale of attention is relative to each base model and error. We highlight POS tags that are attended to the most across models.

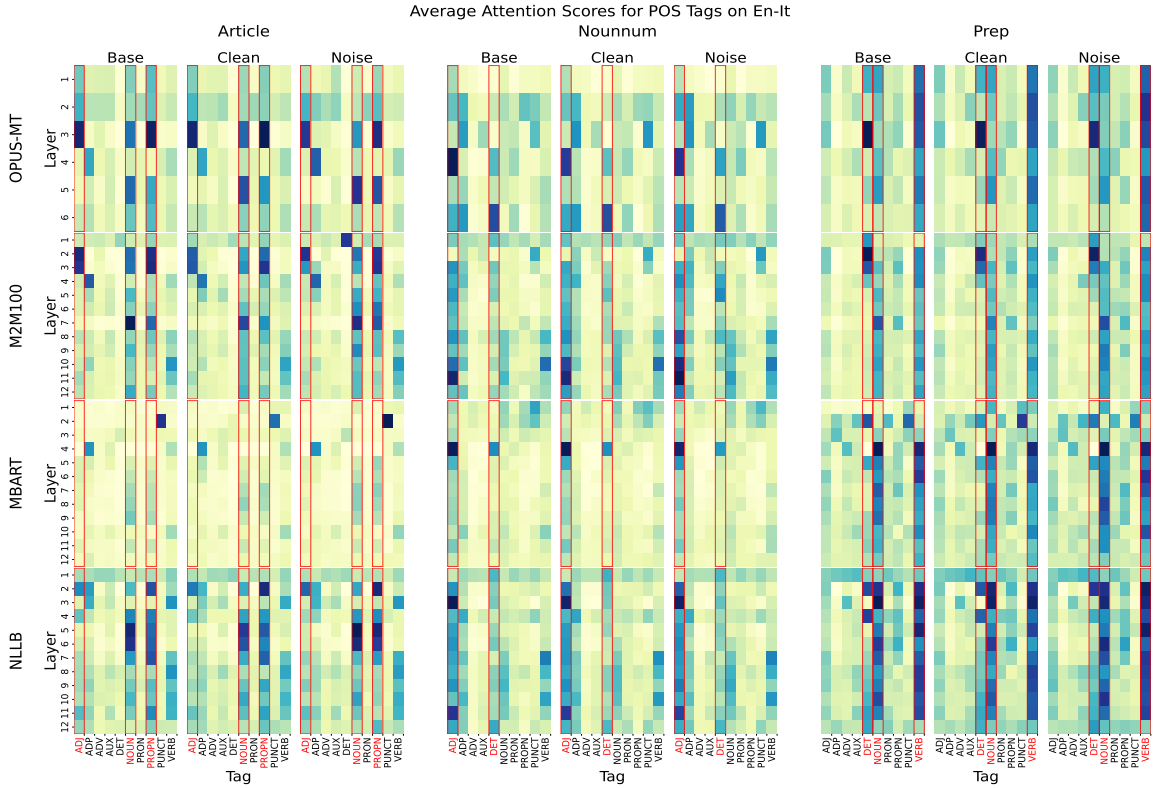


Figure 10: *Robustness Heads* attention to the 10 most common POS tags in the test set on En-It. The scale of attention is relative to each base model and error. We highlight POS tags that are attended to the most across models.

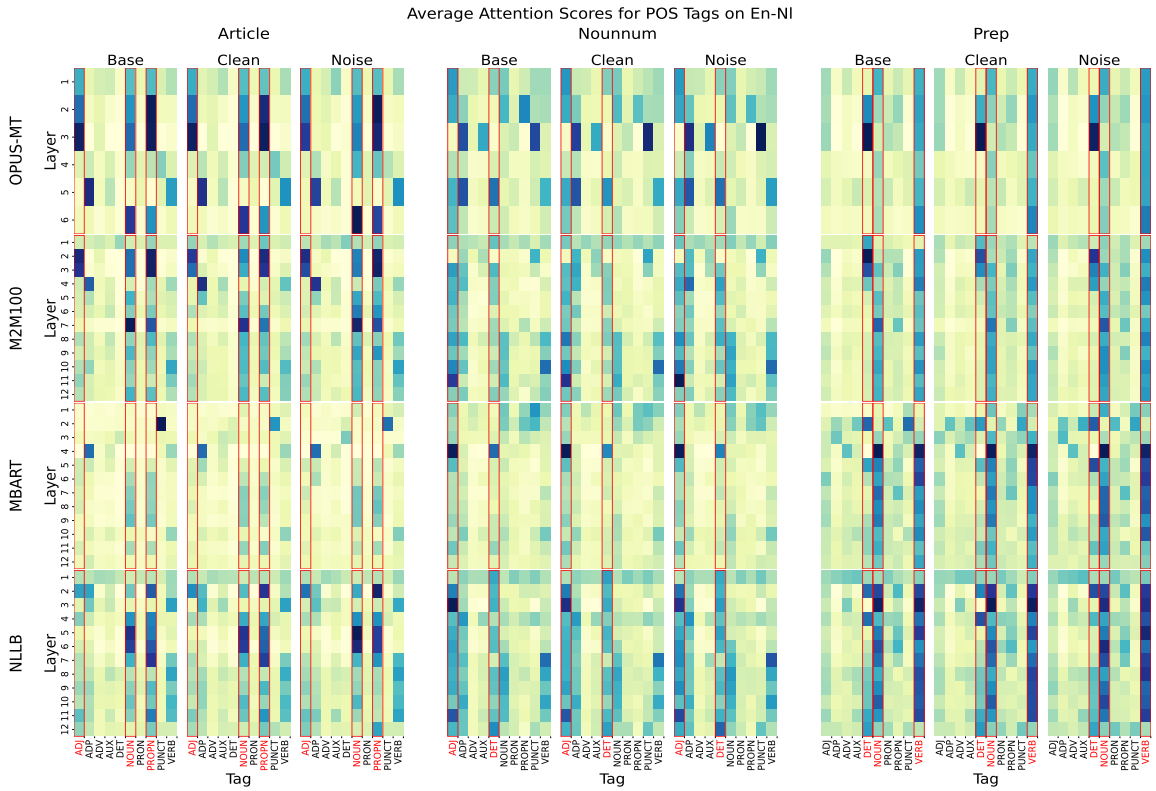
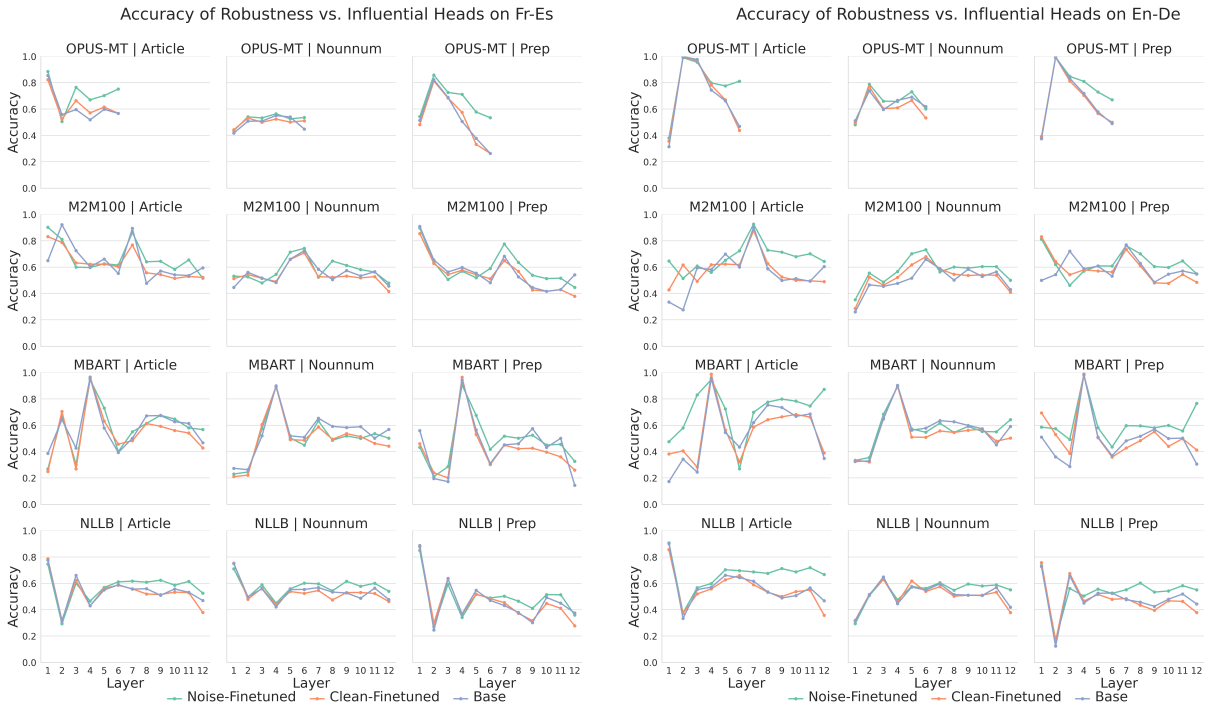
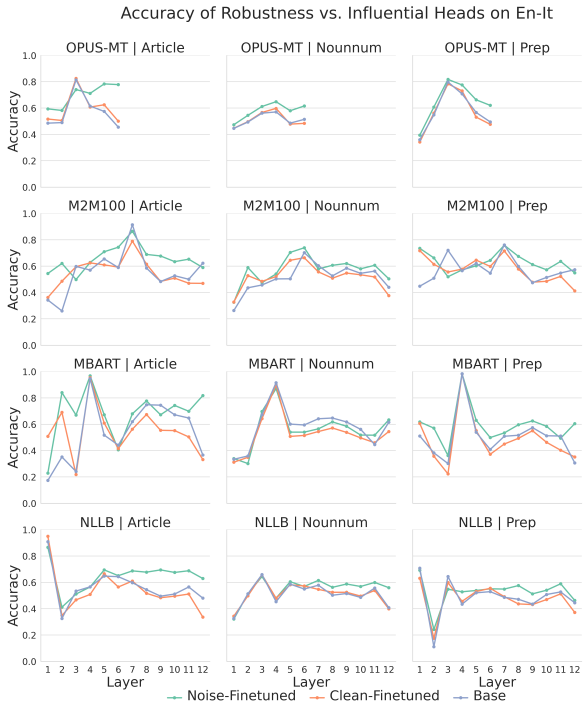


Figure 11: *Robustness Heads* attention to the 10 most common POS tags in the test set on En-Nl. The scale of attention is relative to each base model and error. We highlight POS tags that are attended to the most across models.

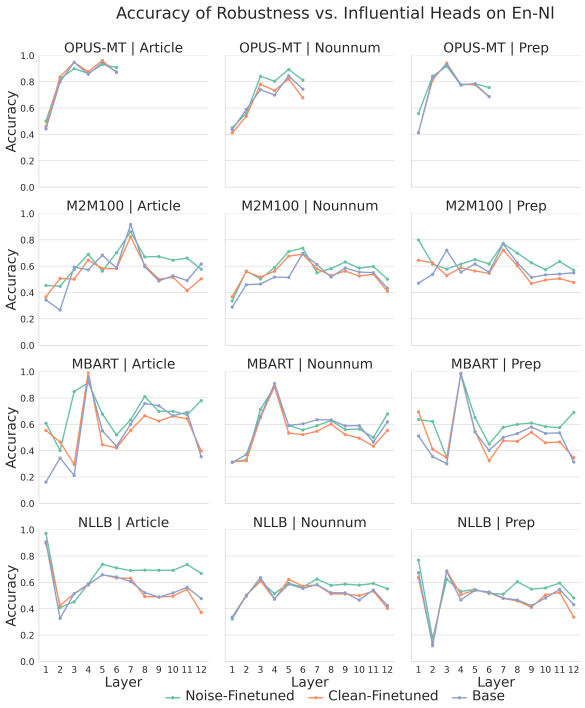


(a) Fr-Es

(b) En-De



(c) En-It



(d) En-Nl

Figure 12: Accuracy of Robustness and Influential heads on Fr-Es, En-De, En-It and En-Nl. We find the accuracy is higher for Noise-Finetuned models especially in deep layers.

Direction	Model	Morpheus Clean	Morpheus Noisy	Morpheus Δ
En-Es	opus-base	76.03	71.78	4.25
	opus-clean	**76.2**	72.28	3.92
	opus-noisy	75.85	**75.35**	**0.5**
	m2m100-base	75.98	71.8	4.18
	m2m100-clean	**76.48**	72.54	3.94
	m2m100-noisy	75.79	**75.58**	**0.22**
	mbart-base	76.44	72.75	3.69
	mbart-clean	**78.38**	74.77	3.62
	mbart-noisy	78.27	**78.08**	**0.19**
	nllb-base	75.18	71.78	3.4
	nllb-clean	**75.83**	72.46	3.36
	nllb-noisy	75.64	**75.45**	**0.19**
Fr-Es	opus-base	71.18	67.63	3.55
	opus-clean	**73.97**	70.71	3.26
	opus-noisy	73.57	**73.05**	**0.52**
	m2m100-base	73.0	68.64	4.35
	m2m100-clean	73.3	70.2	3.09
	m2m100-noisy	**73.42**	**73.23**	**0.19**
	mbart-base	65.36	60.31	5.06
	mbart-clean	73.07	69.69	3.39
	mbart-noisy	**74.12**	**73.62**	**0.5**
	nllb-base	71.34	68.53	2.81
	nllb-clean	73.1	70.68	2.42
	nllb-noisy	**73.74**	**73.43**	**0.3**
En-De	opus-base	68.62	63.91	4.71
	opus-clean	**70.1**	65.29	4.81
	opus-noisy	69.15	**68.4**	**0.75**
	m2m100-base	72.92	69.02	3.9
	m2m100-clean	**74.53**	71.3	3.23
	m2m100-noisy	73.98	**73.65**	**0.33**
	mbart-base	74.52	69.94	4.57
	mbart-clean	**77.38**	74.1	3.28
	mbart-noisy	77.02	**76.82**	**0.2**
	nllb-base	65.19	62.09	3.1
	nllb-clean	66.24	63.45	2.78
	nllb-noisy	**66.52**	**66.02**	**0.5**
En-It	opus-base	74.82	70.14	4.69
	opus-clean	76.16	72.11	4.05
	opus-noisy	**76.22**	**75.73**	**0.48**
	m2m100-base	74.94	70.97	3.98
	m2m100-clean	**76.69**	73.38	3.31
	m2m100-noisy	76.59	**76.46**	**0.13**
	mbart-base	75.22	72.01	3.22
	mbart-clean	77.55	74.2	3.35
	mbart-noisy	**77.58**	**77.37**	**0.22**
	nllb-base	74.84	71.68	3.17
	nllb-clean	**75.78**	72.82	2.96
	nllb-noisy	75.59	**75.26**	**0.33**
En-Nl	opus-base	74.07	69.39	4.68
	opus-clean	75.32	71.5	3.82
	opus-noisy	**76.08**	**75.51**	**0.58**
	m2m100-base	74.84	71.45	3.4
	m2m100-clean	**75.28**	71.99	3.29
	m2m100-noisy	74.78	**74.63**	**0.15**
	mbart-base	74.09	71.46	2.63
	mbart-clean	77.25	74.4	2.85
	mbart-noisy	**77.38**	**77.28**	**0.1**
	nllb-base	72.61	69.73	2.87
	nllb-clean	**73.64**	71.13	2.51
	nllb-noisy	73.53	**73.39**	**0.15**

Table 4: COMET scores for Morpheus on En-Es, Fr-Es, En-De, En-It and En-NL. The Base model is the original model, Clean is fine-tuned on the clean version of the data, and Noise is fine-tuned on the noisy version (with the same noise as the one they are tested on). We present the performance on the clean and noisy test sets and their difference (Δ).

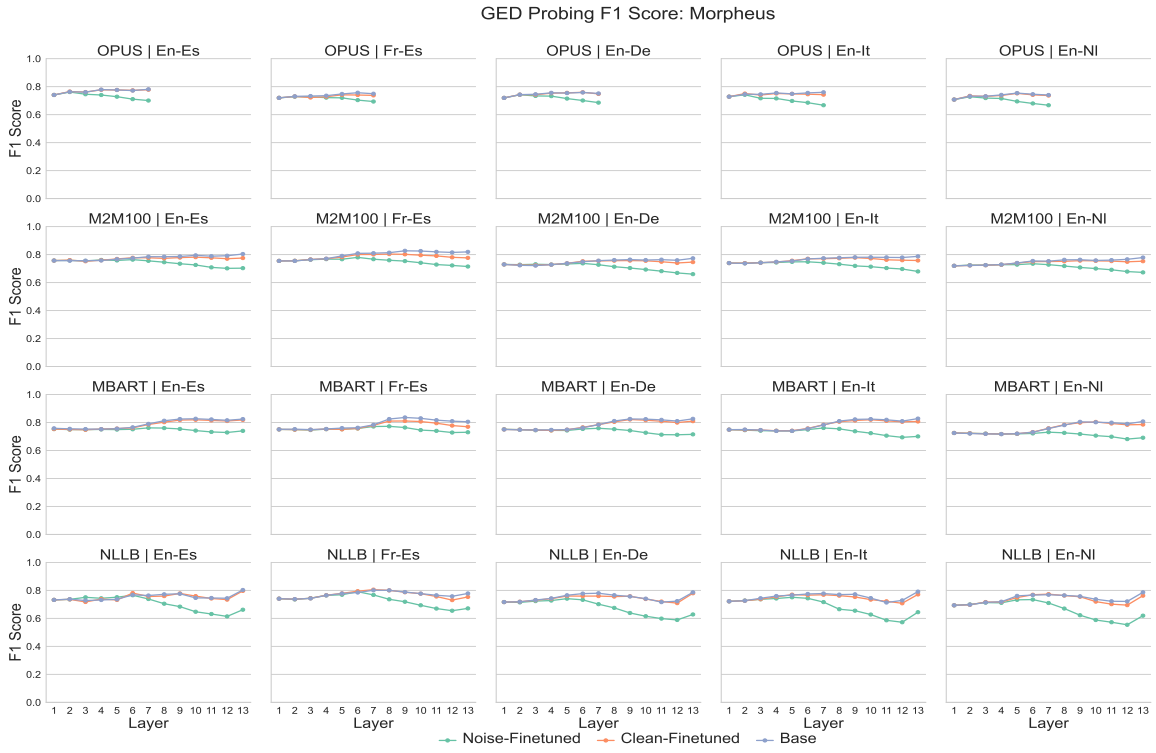


Figure 13: GED probing performance of Noise-Finetuned, Clean-Finetuned and Base models on En-Es, Fr-Es, En-De, En-It and En-Nl on Morpheus errors. GED probing performance of Noise-Finetuned models witnesses a degradation in deeper layers.

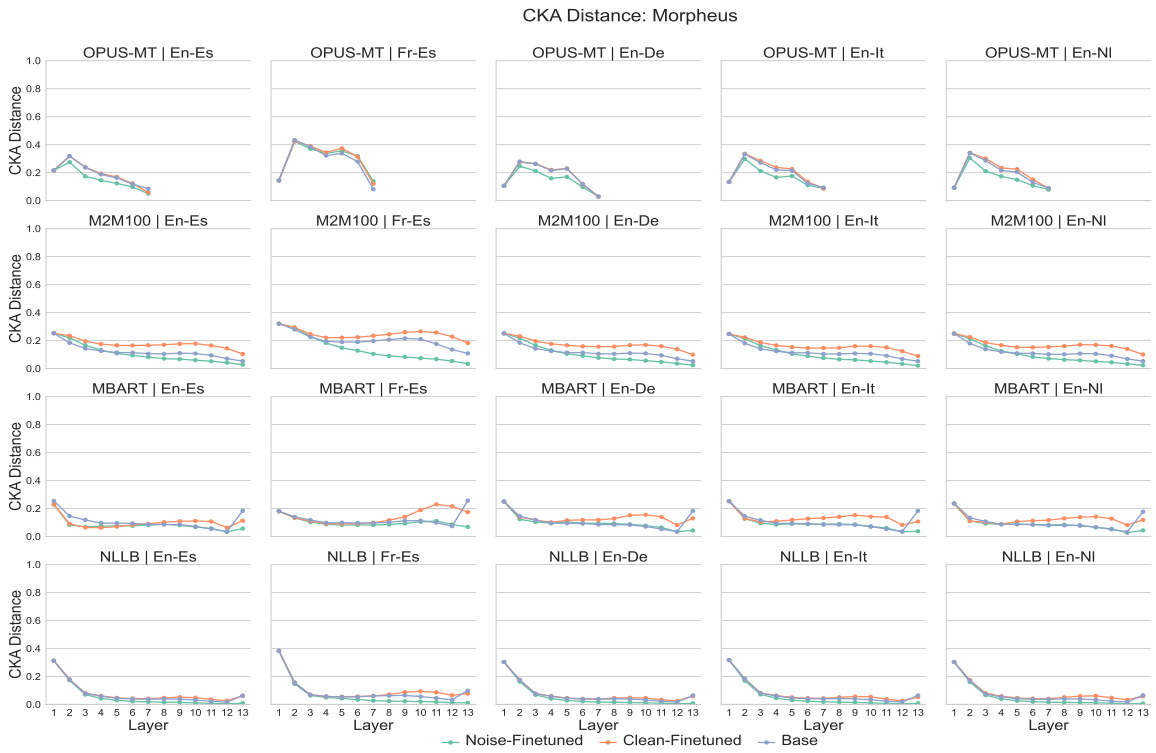


Figure 14: CKA distance of clean and noise word representations across models and errors on En-Es, Fr-Es, En-De, En-It and En-Nl on Morpheus Errors. Noise-Finetuned models drive the representations of the noisy word to be more similar to the clean word.

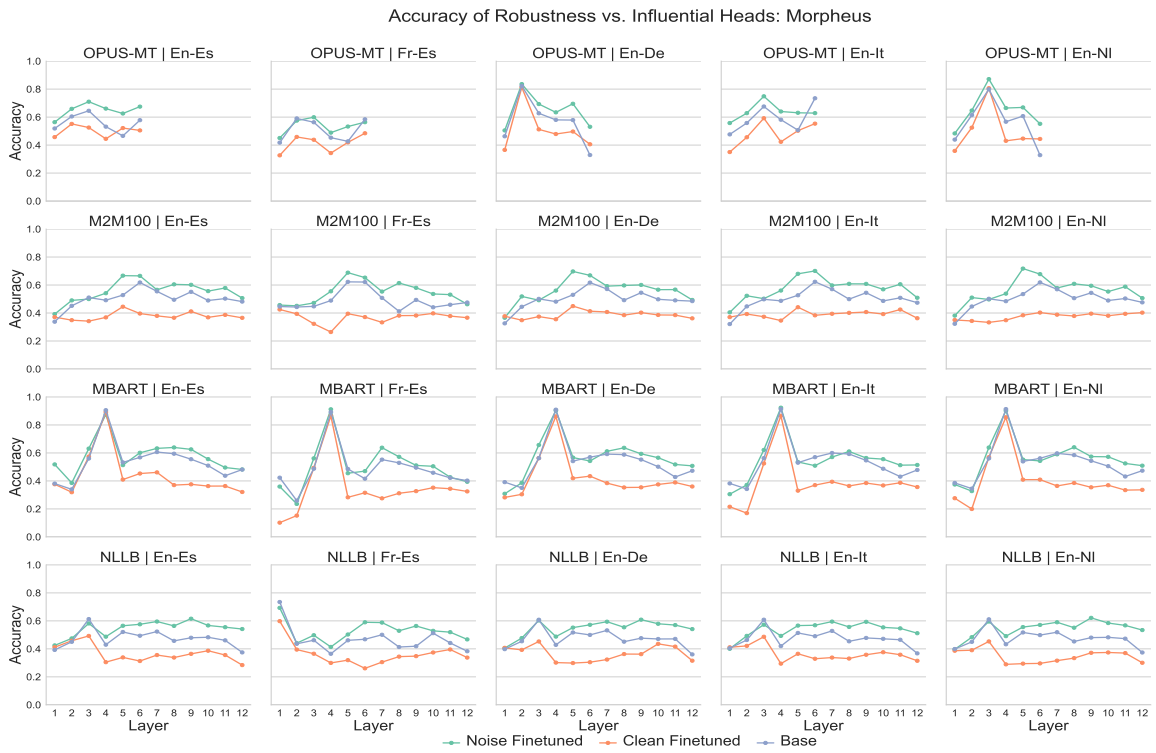


Figure 15: Accuracy of Robustness and Influential heads on En-Es, Fr-Es, En-De, En-It and En-Nl on Morpheus errors. We find the accuracy is higher for Noise-Finetuned models especially in deep layers.