Embed_Llama: using LLM embeddings for the Metrics Shared Task

Sören Dréano
ML-Labs
Dublin City University
soren.dreano2@mail.dcu.ie

Derek Molloy
School of Electronic Engineering
Dublin City University
derek.molloy@dcu.ie

Noel Murphy
School of Electronic Engineering
Dublin City University
noel.murphy@dcu.ie

Abstract

Embed_llama is an assessment metric for language translation that hinges upon the utilization of the recently introduced Llama 2 Large Language Model (LLM), specifically focusing on its embedding layer, to transform sentences into a vector space that establishes connections between geometric and semantic proximities.

Investigations utilizing previous WMT datasets have revealed that within the Llama 2 architecture, relying solely on the initial embedding layer does not result in the highest degree of correlation when assessing machine translations. The incorporation of additional layers, however, holds the potential to augment the contextual understanding of sentences.

As a contribution to the WMT23 challenge, this study delves into the advantages derived from employing a pre-trained LLM that has not undergone fine-tuning specifically for translation evaluation tasks, to provide a metric conducive to operation on readily accessible consumer-grade hardware. This research digs into the observation that deeper layers within the model do not result in a linear increase in the spatial proximity between sentences within the vector space.

1 Introduction

The assessment of algorithm-generated translations entails the utilization of evaluation metrics that furnish quantitative scores to objectively gauge the precision of the model's output. Various methodologies are employed to juxtapose machine-generated translations with their human-generated counterparts, contingent upon the specific evaluation metric employed. In recent years, the realm of machine translation (MT) has witnessed notable advancements in terms of both translation accuracy and linguistic fluency.

Over time, there has been a significant enhancement in the correlation coefficient between human assessments and the automated evaluation of sentences generated by machines. Earlier metrics, such as BLEU (Papineni et al., 2002) or chrF++ (Popović, 2017), predominantly relied on the textual overlap between reference translations and the machine-generated counterparts. In contrast, contemporary approaches, exemplified by COMET (Rei et al., 2020), harness recent breakthroughs in Natural Language Processing(NLP) and transformer models, enabling them to consider not only individual words, but also to leverage contextual semantics for a more comprehensive evaluation.

In the domain of Machine Learning (ML) applied to NLP, the embedding layer assumes a pivotal role within neural network architectures, particularly in tasks centered on textual data. Its central objective lies in the transformation of discrete tokens, encompassing entities like words or characters into continuous vector representations. These vector representations, which maintain continuity, are amenable to acquisition and manipulation by neural networks and are commonly referred to as *word embeddings*.

2 Embed_Llama

The initial component in a Natural Language Processing (NLP) model is typically an embedding layer, which serves the purpose of converting the distinct identifiers of tokens within the input sentence into a vectorized representation. In this context, it is essential to emphasize that sentences conveying similar semantic content should exhibit proximity in the vector space, irrespective of the presence of word-level overlap, in contrast to sentences chosen randomly.

Embed_Llama draws inspiration from Word2vec (Mikolov et al., 2013) using Llama 2 (Touvron et al., 2023), a contemporary open-source pre-trained model. Rather than needing to train an extra NLP model for the purpose of assessing translation quality, a viable alternative approach involves uti-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 738–745
December 6–7, 2023.

lizing a pre-trained, extensive neural network like Llama 2, which has been originally trained for next-token prediction. This approach allows for the investigation of how closely related sentences evolve across the model's various layers, all without incurring the supplementary expenses associated with fine-tuning or training anew.

2.1 Word2vec

Word2vec is a methodology that utilizes a neural network model to extract word associations from comprehensive textual corpora. Post training, this model possesses the capability to identify synonymous terms and offer word suggestions for unfinished sentences. As the terminology implies, Word2vec symbolizes individual words by employing distinct numerical vectors, systematically engineered to encapsulate both the semantic and syntactic characteristics inherent to the words.

Embed_Llama leverages vectorial space to estimate similarity or dissimilarity. This estimation is accomplished by computing the cosine distance between two sentences.

2.2 Vocabulary size

The lexical repertoire of Llama 2 encompasses 32,000 unique tokens, a figure lower than that of both GPT-2 (Radford et al., 2019) and GPT-NeoX (Black et al., 2022), which both employ 50,000 unique tokens. Regrettably, the specific vocabulary sizes pertaining to GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) remain undisclosed.

2.3 Embeddings

The dimensionality of the embedding space represents a hyperparameter subject to adjustment. Embeddings of higher dimensions have the capacity to capture more intricate relationships; however, it is noteworthy that such higher-dimensional embeddings may necessitate increased quantities of data and computational resources for their effective utilization.

2.4 Cosine distance

The cosine similarity metric quantifies the similarity between two n-dimensional vectors by computing the cosine of the angle between them. This scoring measure finds common application in the domain of text mining (Singhal, 2001). The general formula for two vectors A and B is:

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

To enable efficient computation on consumer-grade hardware, the two sentences slated for comparison are padded to match the maximum token count of the longer sentence in the pair. Consequently, a sentence initially shaped as [length] transforms into the shape [length ×4,096] following its processing through the embedding layer.

3 Hyperparameter

As the objective of this current study revolves around the utilization of a pre-trained network, our sphere of influence is limited to a sole hyperparameter, namely, the number of blocks to retain before computing the cosine distance.

3.1 Block architecture

As shown in Figure 1, each block, denoted as the *LlamaDecoderLayer*, is structured with several components, including an attention layer, two normalization layers, and a multi-layer perceptron. The multi-layer perceptron, in turn, consists of three linear layers along with an associated activation function, while the attention layer also includes a rotary embedding layer.

The Llama 2 model, comprising 7 billion parameters, encompasses an embedding layer, 32 blocks, and a projection layer. To determine the optimal number of blocks, datasets extracted from the WMT challenge editions of 2020, 2021, and 2022 were employed. Due to the limited GPU memory allocation in the current project, it was only feasible to investigate the Llama 2 model up to a depth of 22 blocks, whereas the model has a total of 32 available blocks.

```
LlamaDecoderLayer(
  (self_attn): LlamaAttention(
    (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (k_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (v_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (mlp): LlamaMLP(
    (gate_proj): Linear(in_features=4096, out_features=11008, bias=False)
    (up_proj): Linear(in_features=4096, out_features=11008, bias=False)
    (down_proj): Linear(in_features=11008, out_features=4096, bias=False)
    (act_fn): SiLUActivation()
  )
  (input_layernorm): LlamaRMSNorm()
  (post_attention_layernorm): LlamaRMSNorm()
)
```

Figure 1: Architecture of the Llama2 model as displayed in the Huggingface library

3.2 Exploring the depth

It was initially hypothesized that increasing the number of blocks would improve the contextual

representation of sentence meaning. Figure 2 reveals that, in contrast to our initial hypotheses, the augmentation of block quantities does not significantly modify the correlation between the systems and the ground truth, whether by augmentation or reduction. Furthermore, it is noteworthy that this correlation exhibits variability across different datasets. Specifically, the Pearson correlation between the number of layers and algorithm accuracy is 0.34 for the WMT20 dataset, but it decreases to -0.79 for the WMT22 dataset.

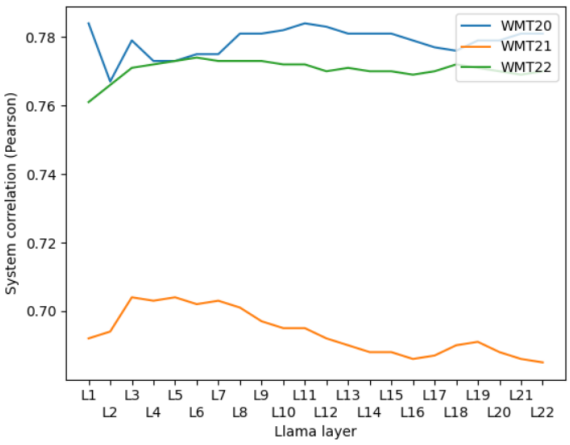

Figure 2: System correlations depending on the Llama layer

As shown in Figure 3, these observations hold with respect to segment correlation as well. The quantity of blocks employed in the Embed_Llama does not consistently enhance the metric's quality, whether for individual segments or entire systems.

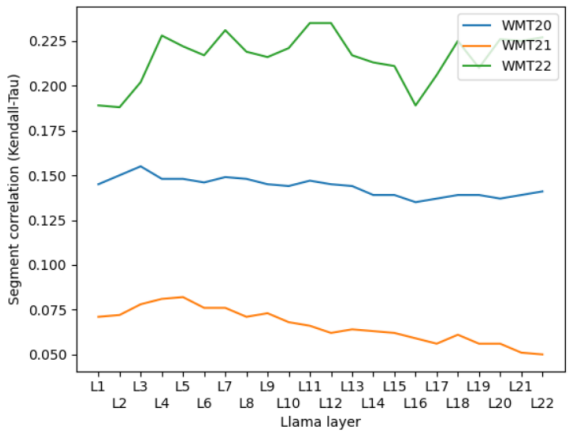

Figure 3: Segment correlations depending on the Llama layer

The highest levels of correlation between human rankings and Embed_Llama rankings were achieved by utilizing a mere two blocks following the embedding layer, resulting in optimal overall performance across the WMT20, WMT21, and WMT22 datasets. This not only expedited the computational process, but also decreased the GPU memory demands for metric computation.

3.3 Inter-languages variations

As depicted in Figure 4, the associations between metric quality and language pairs exhibit large variations. For instance, when considering the Hungarian-to-English language pair, the Pearson coefficient registers at 0.83, whereas it falls to -0.85 for the Czech-to-Ukrainian pair.

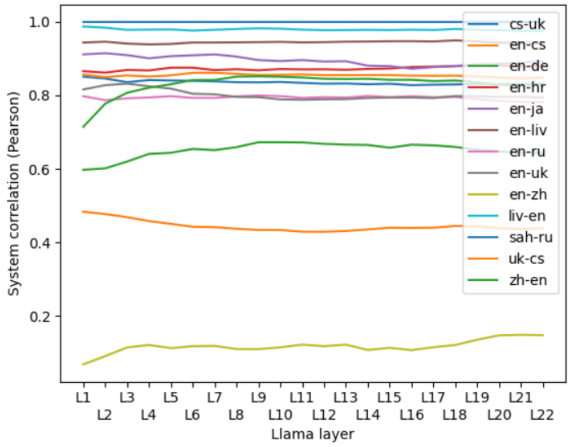

Figure 4: System correlations for each language pair in the WMT2022 dataset depending on the Llama layer

3.4 Intra-languages variations

Considerable variability is observed even within the same language pair across various datasets. For instance, the English-to-Chinese language pair is encompassed within the WMT2020, WMT2021, and WMT2022 datasets. However, as illustrated in Figure 5, no discernible correlations emerge between the number of utilized blocks and the quality of Embed_Llama scores. This is evident in the transformation of the Pearson coefficient, which shifts from -0.28 in the WMT2021 to 0.73 in the WMT2022 dataset.

3.5 Inter-datasets variations

Table 1 presents the average mean values and their corresponding standard deviations, showcasing the relationship between metric accuracy and the number of utilized blocks for language pairs common to all three datasets. It is notable that, apart for the English-to-Japanese pair, there exists a significant degree of variability in the performance of the

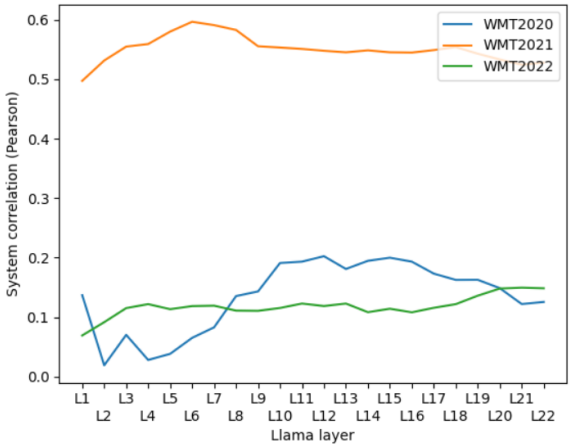

Figure 5: System correlations of the English-to-Chinese language pair depending on the Llama layer and the dataset

Language pair	Mean	Standard deviation
en-cs	-0.34	0.22
en-de	-0.02	0.4
en-ja	-0.89	0.04
en-ru	-0.02	0.54
en-zh	0.35	0.45
zh-en	-0.17	0.46

Table 1: Means and standard deviations of the system correlations depending on the Llama layer when WMT2020, WMT2021 and WMT2022 are merged

same language pairs across different datasets. This variability is underscored by the substantial standard deviations in relation to the absolute values of the means. Given the limited usage of just three datasets, it is essential to acknowledge that the relatively small sample size may hinder the ability to draw conclusive inferences regarding inter-dataset variability.

3.6 Full results

Tables 2, 4 and 6, correspondingly, present the Pearson correlation coefficients for datasets WMT2020, WMT2021 and WMT2021, illustrating the association between the algorithm-assigned scores and the actual rankings of the evaluated systems for individual language pairs.

With regard to segment-level correlations, they are presented in Tables 3, 5 and 7 for WMT2020, WMT2021, and WMT2022, respectively. It is noteworthy that, in contrast to system correlations, these are represented by Kendall coefficients, which are utilized as a measure of ordinal association.

It is noteworthy that the observed variations in these correlations are predominantly influenced by the specific language pairs, rather than the depth of the final block employed prior to cosine similarity computation, aligning with our anticipated outcome.

3.7 Source code

The source code of Embed_Llama is available at https://github.com/SorenDreano/embed_llama.

4 Conclusion

Although the authors initially anticipated that Embed_Llama would exhibit suboptimal performance for a majority of language pairs, except for those involving English, due to the apparent constraints posed by a limited vocabulary size and the nature of the dataset Llama 2 was trained on, the actual performance did not exhibit a significant underperformance.

The results from previous iterations of the WMT metrics shared task, presented in Appendix A, indicate that this approach may not meet the contemporary state-of-the-art standards exemplified by METRICX_XXL (unpublished) and COMET-22 (Rei et al., 2022).

The methodology involving the utilization of a non-finetuned, pre-trained Large Language Model (LLM) to assess translation quality through vector space similarity comparisons remains a prospective avenue of inquiry. This prospect gains relevance in light of forthcoming open-source models characterized by expansive vocabularies and training data encompassing diverse languages.

5 Further work

Given the recent proliferation of open-source LLMs, it is likely that another model, either presently or in the near future, may surpass the performance of Llama 2 for translation evaluation without necessitating any fine-tuning.

In the current investigation, the exploration has been confined to the 7 billion parameters model. It remains conceivable that employing a more extensive model with increased parameters may yield a more precise metric, albeit at the trade-off of heightened computational resource demands.

In the present evaluation, an exploration was limited to the initial 22 blocks. Subsequent endeavors

may consider augmenting this number, as doing so could potentially result in further benefits.

Moreover, it is worth noting that the optimal selection of the number of blocks to employ may be contingent upon the specific target language. Consequently, adjusting this hyperparameter based on the language in question could potentially yield enhanced correlation results.

In the scope of the current academic study, solely the cosine distance served as the chosen similarity measure for tensors. Future research endeavors may wish to investigate alternative distance metrics, such as the Euclidean distance or the Manhattan distance, for potential exploration.

6 Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

OpenAI. 2023. Gpt-4 technical report.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24:35–43.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

A Appendix. Tables

Table 2: Pearson correlations (r) for the WMT20 system dataset depending on the Llama layer. The findings pertaining to the second layer are shown in bold, as it represents the prevailing default layer count within the Embed_Llama

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.855	**0.866**	0.875	0.877	0.873	0.876	0.872	0.863	0.861	0.864	0.867
en-de	0.923	**0.928**	0.933	0.929	0.928	0.927	0.927	0.926	0.925	0.925	0.928
en-ru	0.856	**0.677**	0.769	0.772	0.81	0.829	0.821	0.902	0.912	0.899	0.918
en-ta	0.883	**0.901**	0.914	0.921	0.926	0.932	0.934	0.936	0.939	0.942	0.945
en-zh	0.137	**0.019**	0.07	0.028	0.038	0.065	0.083	0.135	0.143	0.191	0.193
en-ja	0.898	**0.892**	0.879	0.871	0.88	0.881	0.874	0.883	0.88	0.876	0.871
en-pl	0.88	**0.874**	0.88	0.878	0.875	0.864	0.868	0.873	0.873	0.872	0.868
en-iu	0.252	**0.226**	0.194	0.156	0.137	0.121	0.119	0.106	0.105	0.096	0.094
cs-en	0.795	**0.777**	0.752	0.771	0.789	0.788	0.796	0.795	0.796	0.789	0.775
de-en	0.992	**0.996**	0.996	0.994	0.99	0.99	0.99	0.988	0.988	0.988	0.99
pl-en	0.419	**0.403**	0.428	0.433	0.401	0.395	0.4	0.394	0.388	0.398	0.407
ta-en	0.876	**0.889**	0.918	0.922	0.919	0.923	0.921	0.919	0.917	0.919	0.918
km-en	0.954	**0.969**	0.988	0.988	0.988	0.989	0.99	0.989	0.987	0.986	0.985
ps-en	0.926	**0.874**	0.877	0.862	0.872	0.877	0.873	0.875	0.875	0.875	0.878
ja-en	0.913	**0.938**	0.947	0.946	0.946	0.948	0.949	0.948	0.946	0.948	0.949
ru-en	0.949	**0.939**	0.949	0.948	0.953	0.947	0.948	0.949	0.948	0.944	0.939
zh-en	0.97	**0.967**	0.963	0.956	0.957	0.955	0.955	0.954	0.954	0.951	0.951
iu-en	0.64	**0.676**	0.68	0.662	0.638	0.645	0.638	0.625	0.616	0.615	0.633

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.865	0.866	0.86	0.858	0.851	0.856	0.855	0.857	0.866	0.874	0.874
en-de	0.925	0.926	0.922	0.921	0.917	0.918	0.918	0.921	0.923	0.929	0.929
en-ru	0.922	0.904	0.903	0.898	0.902	0.885	0.862	0.902	0.93	0.941	0.938
en-ta	0.942	0.942	0.944	0.946	0.947	0.947	0.948	0.948	0.948	0.946	0.946
en-zh	0.202	0.181	0.195	0.2	0.193	0.173	0.162	0.163	0.149	0.122	0.125
en-ja	0.87	0.867	0.869	0.865	0.865	0.857	0.858	0.848	0.838	0.836	0.836
en-pl	0.873	0.871	0.875	0.872	0.869	0.86	0.865	0.862	0.858	0.857	0.857
en-iu	0.084	0.085	0.082	0.084	0.083	0.078	0.073	0.082	0.091	0.092	0.086
cs-en	0.781	0.787	0.787	0.794	0.802	0.794	0.8	0.79	0.783	0.776	0.785
de-en	0.989	0.989	0.988	0.988	0.986	0.989	0.989	0.991	0.992	0.994	0.994
pl-en	0.408	0.42	0.421	0.418	0.417	0.414	0.413	0.41	0.406	0.415	0.415
ta-en	0.918	0.917	0.914	0.91	0.907	0.912	0.912	0.913	0.913	0.919	0.921
km-en	0.983	0.982	0.975	0.975	0.967	0.964	0.967	0.968	0.97	0.966	0.968
ps-en	0.876	0.881	0.887	0.892	0.891	0.888	0.889	0.889	0.889	0.895	0.897
ja-en	0.947	0.946	0.944	0.941	0.936	0.94	0.942	0.943	0.942	0.941	0.941
ru-en	0.938	0.938	0.938	0.939	0.94	0.935	0.935	0.935	0.933	0.935	0.937
zh-en	0.95	0.951	0.951	0.95	0.952	0.953	0.951	0.954	0.955	0.958	0.958
iu-en	0.619	0.613	0.606	0.6	0.602	0.627	0.63	0.639	0.644	0.662	0.66

Table 3: Kendall correlations (τ) for the WMT20 segment dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.226	**0.231**	0.244	0.234	0.232	0.226	0.229	0.225	0.225	0.222	0.226
en-de	0.181	**0.19**	0.207	0.201	0.205	0.198	0.202	0.201	0.199	0.196	0.198
en-ru	0.028	**0.034**	0.037	0.045	0.038	0.045	0.042	0.044	0.043	0.046	0.05
en-ta	0.355	**0.354**	0.323	0.284	0.282	0.267	0.279	0.281	0.282	0.26	0.255
en-zh	0.147	**0.141**	0.164	0.144	0.138	0.141	0.147	0.152	0.152	0.16	0.165
en-ja	0.277	**0.281**	0.295	0.29	0.285	0.287	0.283	0.285	0.284	0.284	0.282
en-pl	0.097	**0.095**	0.103	0.101	0.101	0.102	0.098	0.101	0.098	0.094	0.102
en-iu	0.218	**0.224**	0.205	0.187	0.18	0.173	0.185	0.187	0.186	0.179	0.18
cs-en	0.064	**0.065**	0.073	0.072	0.077	0.076	0.076	0.073	0.074	0.078	0.083
de-en	0.349	**0.374**	0.387	0.384	0.39	0.389	0.385	0.376	0.372	0.375	0.381
pl-en	-0.016	**-0.016**	-0.0	-0.007	-0.0	0.0	-0.003	0.002	-0.002	-0.005	-0.006
ta-en	0.108	**0.118**	0.125	0.121	0.122	0.119	0.128	0.121	0.106	0.108	0.123
km-en	0.141	**0.145**	0.144	0.12	0.13	0.122	0.117	0.114	0.097	0.108	0.114
ps-en	0.088	**0.081**	0.074	0.076	0.058	0.053	0.06	0.061	0.065	0.061	0.08
ja-en	0.109	**0.136**	0.136	0.144	0.147	0.149	0.151	0.144	0.137	0.139	0.134
ru-en	0.011	**0.02**	0.019	0.022	0.026	0.022	0.031	0.023	0.02	0.02	0.017
zh-en	0.065	**0.067**	0.071	0.072	0.073	0.072	0.072	0.071	0.068	0.069	0.069
iu-en	0.16	**0.161**	0.175	0.178	0.187	0.191	0.195	0.194	0.195	0.198	0.199

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.23	0.229	0.222	0.223	0.215	0.218	0.221	0.222	0.221	0.222	0.221
en-de	0.19	0.193	0.187	0.183	0.177	0.178	0.184	0.184	0.186	0.186	0.19
en-ru	0.043	0.037	0.043	0.042	0.045	0.036	0.039	0.037	0.039	0.043	0.044
en-ta	0.281	0.29	0.283	0.266	0.261	0.267	0.264	0.232	0.204	0.224	0.237
en-zh	0.166	0.168	0.162	0.165	0.16	0.163	0.166	0.163	0.165	0.166	0.164
en-ja	0.282	0.276	0.28	0.281	0.274	0.272	0.271	0.271	0.263	0.266	0.27
en-pl	0.098	0.096	0.091	0.087	0.093	0.093	0.095	0.093	0.09	0.092	0.092
en-iu	0.17	0.167	0.162	0.163	0.154	0.153	0.149	0.146	0.141	0.145	0.143
cs-en	0.081	0.073	0.073	0.076	0.074	0.079	0.081	0.081	0.077	0.073	0.075
de-en	0.377	0.377	0.371	0.372	0.361	0.37	0.37	0.378	0.374	0.383	0.386
pl-en	-0.004	0.0	-0.001	0.005	0.001	-0.001	0.001	-0.004	-0.006	-0.004	-0.003
ta-en	0.118	0.111	0.107	0.101	0.103	0.11	0.112	0.114	0.112	0.106	0.103
km-en	0.109	0.11	0.092	0.097	0.083	0.091	0.1	0.111	0.116	0.124	0.126
ps-en	0.074	0.081	0.067	0.068	0.064	0.056	0.066	0.07	0.07	0.067	0.07
ja-en	0.128	0.125	0.116	0.112	0.102	0.115	0.117	0.125	0.126	0.129	0.129
ru-en	0.015	0.007	0.009	0.008	0.013	0.012	0.015	0.018	0.02	0.025	0.022
zh-en	0.069	0.072	0.07	0.069	0.066	0.069	0.07	0.072	0.07	0.074	0.074
iu-en	0.192	0.177	0.173	0.176	0.173	0.179	0.184	0.183	0.188	0.189	0.189

Table 4: Pearson correlations (r) for the WMT21 system dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.985	**0.984**	0.985	0.982	0.981	0.981	0.982	0.982	0.979	0.98	0.982
en-zh	0.497	**0.531**	0.555	0.559	0.58	0.597	0.591	0.583	0.555	0.553	0.551
en-ha	0.534	**0.551**	0.567	0.567	0.567	0.553	0.554	0.553	0.563	0.56	0.547
en-ja	0.82	**0.83**	0.838	0.836	0.83	0.84	0.837	0.839	0.829	0.822	0.816
en-ru	0.567	**0.621**	0.682	0.674	0.674	0.658	0.674	0.614	0.572	0.576	0.573
en-de	0.818	**0.793**	0.804	0.801	0.794	0.789	0.794	0.799	0.796	0.792	0.79
cs-en	0.542	**0.454**	0.435	0.428	0.429	0.432	0.426	0.426	0.429	0.42	0.423
zh-en	0.232	**0.186**	0.155	0.159	0.172	0.179	0.188	0.196	0.198	0.196	0.192
ha-en	0.825	**0.855**	0.858	0.855	0.867	0.868	0.867	0.867	0.865	0.86	0.855
ja-en	0.728	**0.726**	0.748	0.753	0.761	0.76	0.759	0.759	0.761	0.761	0.759
ru-en	0.613	**0.606**	0.535	0.514	0.5	0.496	0.502	0.506	0.506	0.504	0.509
de-en	0.171	**0.22**	0.237	0.238	0.221	0.23	0.223	0.223	0.225	0.234	0.243
fr-de	0.564	**0.556**	0.555	0.556	0.556	0.555	0.557	0.555	0.556	0.554	0.552
de-fr	0.477	**0.511**	0.578	0.579	0.579	0.58	0.579	0.575	0.563	0.564	0.574
bn-hi	0.908	**0.943**	0.94	0.942	0.937	0.929	0.941	0.95	0.943	0.941	0.942
hi-bn	0.879	**0.872**	0.913	0.93	0.925	0.915	0.912	0.907	0.911	0.908	0.915
xh-zu	0.952	**0.932**	0.934	0.931	0.923	0.909	0.904	0.905	0.898	0.899	0.891
zu-xh	0.95	**0.937**	0.97	0.973	0.975	0.972	0.971	0.972	0.976	0.976	0.976

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.983	0.984	0.983	0.983	0.979	0.981	0.982	0.983	0.983	0.984	0.983
en-zh	0.548	0.545	0.548	0.545	0.545	0.549	0.554	0.543	0.533	0.524	0.528
en-ha	0.541	0.543	0.554	0.553	0.571	0.562	0.572	0.567	0.556	0.542	0.534
en-ja	0.81	0.804	0.795	0.793	0.781	0.783	0.786	0.777	0.772	0.774	0.779
en-ru	0.581	0.582	0.562	0.58	0.526	0.557	0.565	0.603	0.621	0.625	0.637
en-de	0.796	0.803	0.808	0.808	0.814	0.81	0.81	0.8	0.79	0.787	0.791
cs-en	0.394	0.386	0.383	0.387	0.388	0.395	0.41	0.425	0.422	0.433	0.435
zh-en	0.189	0.189	0.185	0.189	0.19	0.173	0.174	0.168	0.157	0.142	0.146
ha-en	0.846	0.842	0.84	0.841	0.841	0.827	0.828	0.829	0.823	0.818	0.814
ja-en	0.76	0.763	0.768	0.771	0.775	0.772	0.772	0.77	0.769	0.762	0.763
ru-en	0.51	0.512	0.513	0.52	0.521	0.52	0.529	0.534	0.535	0.532	0.528
de-en	0.23	0.217	0.212	0.196	0.184	0.207	0.21	0.205	0.203	0.215	0.214
fr-de	0.553	0.554	0.555	0.555	0.553	0.55	0.55	0.547	0.543	0.541	0.542
de-fr	0.566	0.564	0.558	0.558	0.556	0.567	0.564	0.561	0.563	0.574	0.577
bn-hi	0.935	0.932	0.931	0.931	0.93	0.933	0.933	0.938	0.94	0.93	0.922
hi-bn	0.904	0.892	0.876	0.86	0.869	0.863	0.859	0.865	0.864	0.852	0.829
xh-zu	0.9	0.903	0.906	0.903	0.909	0.905	0.908	0.912	0.903	0.898	0.896
zu-xh	0.977	0.979	0.981	0.981	0.981	0.981	0.977	0.977	0.976	0.969	0.962

Table 5: Kendall correlations (τ) for the WMT21 segment dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.197	**0.198**	0.243	0.243	0.238	0.223	0.221	0.219	0.219	0.214	0.22
en-zh	0.028	**0.028**	0.048	0.042	0.042	0.044	0.039	0.041	0.039	0.037	0.037
en-ha	0.073	**0.077**	0.087	0.081	0.077	0.071	0.071	0.068	0.062	0.061	0.062
en-ja	0.181	**0.184**	0.204	0.21	0.214	0.204	0.197	0.196	0.204	0.196	0.192
en-ru	0.094	**0.101**	0.099	0.099	0.089	0.09	0.091	0.086	0.081	0.087	0.084
en-de	0.241	**0.31**	0.172	0.172	0.241	0.31	0.31	0.241	0.241	0.241	0.241
cs-en	-0.042	**-0.048**	-0.049	-0.045	-0.041	-0.044	-0.042	-0.05	-0.046	-0.045	-0.051
zh-en	0.319	**0.286**	0.319	0.384	0.395	0.319	0.308	0.319	0.384	0.297	0.276
ha-en	-0.022	**-0.016**	-0.007	-0.003	-0.003	-0.009	-0.008	-0.007	-0.011	-0.014	-0.011
ja-en	-0.028	**-0.021**	-0.019	-0.018	-0.015	-0.017	-0.014	-0.015	-0.014	-0.012	-0.015
ru-en	-0.121	**-0.124**	-0.12	-0.12	-0.117	-0.123	-0.12	-0.119	-0.121	-0.124	-0.122
de-en	-0.155	**-0.158**	-0.151	-0.148	-0.153	-0.157	-0.155	-0.156	-0.158	-0.155	-0.152
fr-de	0.057	**0.058**	0.055	0.052	0.044	0.055	0.062	0.057	0.053	0.047	0.051
de-fr	0.054	**0.066**	0.082	0.079	0.075	0.051	0.057	0.054	0.049	0.055	0.055
bn-hi	-0.006	**0.007**	0.016	0.026	0.026	0.024	0.024	0.022	0.015	0.023	0.014
hi-bn	0.132	**0.119**	0.126	0.135	0.139	0.137	0.145	0.15	0.147	0.146	0.14
xh-zu	0.128	**0.129**	0.139	0.13	0.118	0.104	0.101	0.101	0.09	0.082	0.065
zu-xh	0.169	**0.139**	0.181	0.181	0.16	0.129	0.122	0.116	0.12	0.118	0.109

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.213	0.218	0.213	0.22	0.198	0.209	0.209	0.196	0.193	0.194	0.197
en-zh	0.042	0.045	0.048	0.048	0.043	0.043	0.043	0.046	0.041	0.04	0.04
en-ha	0.058	0.06	0.061	0.061	0.061	0.06	0.062	0.064	0.068	0.067	0.067
en-ja	0.189	0.187	0.196	0.193	0.202	0.196	0.195	0.19	0.184	0.179	0.179
en-ru	0.088	0.089	0.109	0.107	0.116	0.116	0.109	0.106	0.108	0.104	0.105
en-de	0.172	0.172	0.172	0.172	0.034	0.034	0.034	0.034	0.034	0.034	0.034
cs-en	-0.047	-0.048	-0.045	-0.055	-0.054	-0.06	-0.056	-0.055	-0.053	-0.058	-0.058
zh-en	0.297	0.319	0.265	0.265	0.33	0.286	0.362	0.286	0.297	0.276	0.276
ha-en	-0.014	-0.013	-0.011	-0.014	-0.012	-0.015	-0.014	-0.014	-0.014	-0.014	-0.014
ja-en	-0.016	-0.02	-0.021	-0.02	-0.02	-0.022	-0.022	-0.021	-0.02	-0.021	-0.018
ru-en	-0.121	-0.121	-0.123	-0.123	-0.123	-0.126	-0.124	-0.116	-0.116	-0.12	-0.121
de-en	-0.152	-0.152	-0.152	-0.154	-0.159	-0.154	-0.154	-0.155	-0.156	-0.155	-0.157
fr-de	0.049	0.047	0.043	0.041	0.031	0.034	0.032	0.036	0.032	0.032	0.031
de-fr	0.039	0.04	0.035	0.039	0.049	0.046	0.053	0.046	0.046	0.046	0.052
bn-hi	0.015	0.02	0.022	0.022	0.019	0.02	0.023	0.023	0.027	0.025	0.025
hi-bn	0.123	0.114	0.114	0.113	0.114	0.12	0.12	0.116	0.108	0.091	0.074
xh-zu	0.076	0.083	0.081	0.094	0.096	0.088	0.098	0.089	0.085	0.083	0.07
zu-xh	0.118	0.122	0.14	0.131	0.141	0.138	0.141	0.144	0.137	0.125	0.122

Table 6: Pearson correlations (r) for the WMT22 system dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.484	**0.477**	0.469	0.458	0.451	0.443	0.442	0.437	0.434	0.434	0.429
en-zh	0.069	**0.091**	0.115	0.122	0.113	0.118	0.119	0.111	0.11	0.115	0.123
en-hr	0.866	**0.862**	0.869	0.868	0.875	0.875	0.869	0.871	0.868	0.871	0.87
en-ja	0.911	**0.914**	0.909	0.901	0.906	0.909	0.911	0.905	0.896	0.893	0.896
en-liv	0.943	**0.946**	0.941	0.938	0.94	0.943	0.943	0.944	0.945	0.945	0.944
en-ru	0.798	**0.787**	0.792	0.794	0.798	0.793	0.793	0.797	0.8	0.798	0.793
en-uk	0.816	**0.828**	0.832	0.825	0.818	0.804	0.803	0.796	0.795	0.789	0.788
en-de	0.598	**0.602**	0.62	0.641	0.644	0.654	0.651	0.659	0.673	0.672	0.672
liv-en	0.987	**0.984**	0.978	0.979	0.979	0.976	0.978	0.98	0.982	0.981	0.978
zh-en	0.715	**0.777**	0.807	0.821	0.83	0.841	0.842	0.851	0.852	0.85	0.848
sah-ru	1.0	**1.0**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
uk-cs	0.857	**0.849**	0.854	0.851	0.854	0.861	0.861	0.858	0.855	0.856	0.857
cs-uk	0.851	**0.846**	0.836	0.841	0.84	0.839	0.838	0.836	0.835	0.836	0.834

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.429	0.431	0.436	0.44	0.44	0.44	0.445	0.444	0.439	0.437	0.439
en-zh	0.119	0.123	0.108	0.114	0.108	0.116	0.122	0.136	0.148	0.149	0.148
en-hr	0.871	0.87	0.872	0.873	0.877	0.878	0.879	0.881	0.881	0.879	0.879
en-ja	0.892	0.893	0.881	0.879	0.872	0.877	0.88	0.884	0.885	0.889	0.891
en-liv	0.944	0.946	0.946	0.947	0.947	0.946	0.949	0.947	0.943	0.942	0.942
en-ru	0.795	0.794	0.798	0.795	0.798	0.795	0.796	0.789	0.785	0.782	0.782
en-uk	0.789	0.79	0.793	0.794	0.794	0.793	0.798	0.798	0.796	0.796	0.794
en-de	0.668	0.666	0.665	0.658	0.666	0.664	0.66	0.652	0.648	0.644	0.644
liv-en	0.977	0.977	0.978	0.977	0.978	0.977	0.98	0.978	0.977	0.975	0.975
zh-en	0.845	0.844	0.845	0.842	0.842	0.839	0.84	0.836	0.828	0.826	0.83
sah-ru	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
uk-cs	0.855	0.855	0.855	0.855	0.854	0.853	0.853	0.851	0.848	0.846	0.848
cs-uk	0.832	0.832	0.83	0.831	0.828	0.829	0.83	0.831	0.83	0.83	0.832

Table 7: Kendall correlations (τ) for the WMT22 segment dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.014	**0.014**	0.023	0.024	0.021	0.024	0.021	0.021	0.015	0.009	0.013
en-zh	-0.041	**-0.03**	-0.034	-0.039	-0.035	-0.024	-0.026	-0.027	-0.026	-0.029	-0.026
en-hr	0.119	**0.115**	0.14	0.125	0.123	0.11	0.117	0.118	0.128	0.123	0.122
en-ja	0.097	**0.098**	0.101	0.091	0.085	0.082	0.081	0.082	0.079	0.08	0.075
en-liv	0.362	**0.332**	0.35	0.34	0.33	0.33	0.334	0.31	0.292	0.322	0.332
en-ru	0.262	**0.227**	0.234	0.248	0.23	0.199	0.188	0.227	0.223	0.244	0.234
en-uk	0.081	**0.063**	0.062	0.062	0.066	0.055	0.044	0.041	0.04	0.048	0.052
en-de	0.4	**0.4**	0.4	0.8	0.8	0.8	1.0	0.8	0.8	0.8	1.0
liv-en	0.247	**0.276**	0.311	0.303	0.302	0.302	0.286	0.298	0.291	0.306	0.314
zh-en	0.179	**0.217**	0.241	0.217	0.195	0.225	0.223	0.213	0.193	0.209	0.203
sah-ru	0.458	**0.419**	0.484	0.492	0.478	0.45	0.456	0.47	0.461	0.456	0.444
uk-cs	0.145	**0.159**	0.16	0.153	0.141	0.125	0.132	0.146	0.167	0.156	0.146
cs-uk	0.133	**0.149**	0.159	0.154	0.154	0.148	0.15	0.151	0.149	0.154	0.151

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.008	0.014	0.014	0.011	0.02	0.018	0.017	0.011	0.011	0.008	0.008
en-zh	-0.03	-0.028	-0.034	-0.033	-0.031	-0.033	-0.033	-0.028	-0.029	-0.027	-0.027
en-hr	0.118	0.11	0.119	0.116	0.109	0.101	0.106	0.118	0.127	0.129	0.128
en-ja	0.081	0.083	0.082	0.083	0.081	0.083	0.084	0.075	0.072	0.074	0.074
en-liv	0.344	0.344	0.346	0.34	0.328	0.328	0.346	0.34	0.34	0.344	0.348
en-ru	0.22	0.202	0.174	0.188	0.16	0.167	0.167	0.202	0.192	0.195	0.206
en-uk	0.049	0.047	0.048	0.048	0.044	0.049	0.042	0.036	0.049	0.038	0.034
en-de	1.0	0.8	0.8	0.8	0.6	0.8	1.0	0.8	1.0	1.0	1.0
liv-en	0.313	0.286	0.261	0.26	0.249	0.258	0.269	0.28	0.272	0.277	0.286
zh-en	0.225	0.219	0.195	0.191	0.179	0.185	0.179	0.183	0.211	0.213	0.209
sah-ru	0.441	0.436	0.45	0.439	0.43	0.427	0.447	0.416	0.408	0.399	0.408
uk-cs	0.144	0.15	0.155	0.153	0.149	0.159	0.158	0.156	0.143	0.137	0.137
cs-uk	0.148	0.152	0.154	0.15	0.14	0.14	0.142	0.141	0.139	0.141	0.141