University of Stuttgart
Cluster of Excellence in Data-integrated Simulation Science

University of Stuttgart Visualization Research Center (VISUS)

Visual-Explainable AI: The Use Case of Language Models

Tanja Munz-Körner, Sebastian Künzel, Daniel Weiskopf

Motivation

We address the research problem of **explainable AI** (Artificial Intelligence) in the context of **language models** using **visualization** and **visual analytics**.

Language Models

AI techniques are applied to handle different language-related problems (e.g., translation and natural-language understanding). Language models can create impressive results, but there are still many challenges, and created results may be incorrect. Explainable AI can help understand why certain predictions were made.

Relation to SimTech

We investigate NLP as part of data-integrated simulation science: in the form of supporting cognitive aspects in the creation of a digital human model, which is one of the visionary examples of the Cluster of Excellence “SimTech”.

Goals for Language Models

Open the black box of language models:

- Transparency and interpretability
- Finding problems / debugging
- Improving prediction accuracy

Strategies to Analyze Deep Learning Models

- **Internal states**

  The exploration of a model's internal states while performing a prediction helps see how states change and contribute to the final prediction.

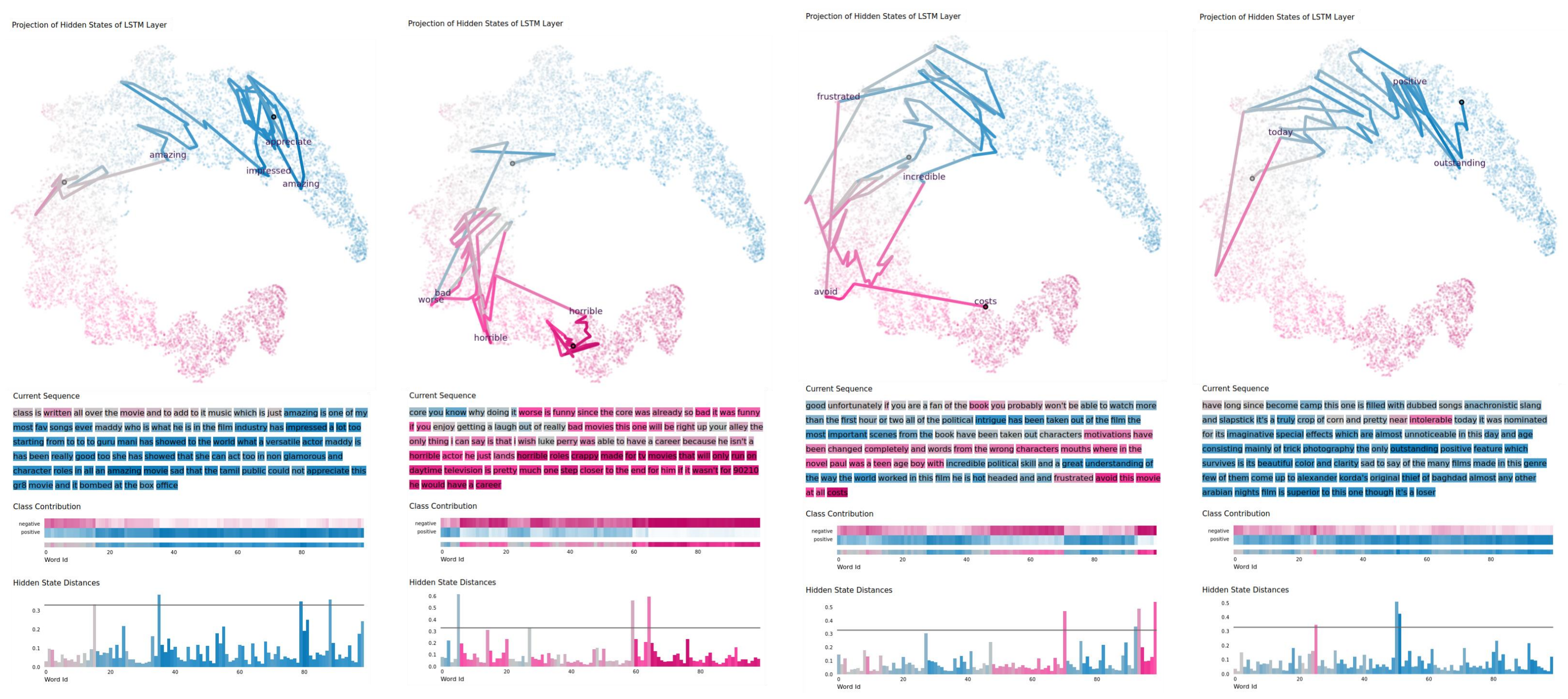
- **Actual prediction result**

  The exploration of the actual prediction results can provide insight into the prediction and provide a quality assessment about the performance of the model.

Approaches

- **Text classification**

Internal structures (hidden states of long short-term memories (LSTMs)) and expected predictions change when processing text sequences and performing a classification task. Our approach [1] shows how the prediction would be if we stopped at an earlier step and how the final prediction was created. For both correct and incorrect predictions, our approach can help analyze what factors influenced such a prediction.



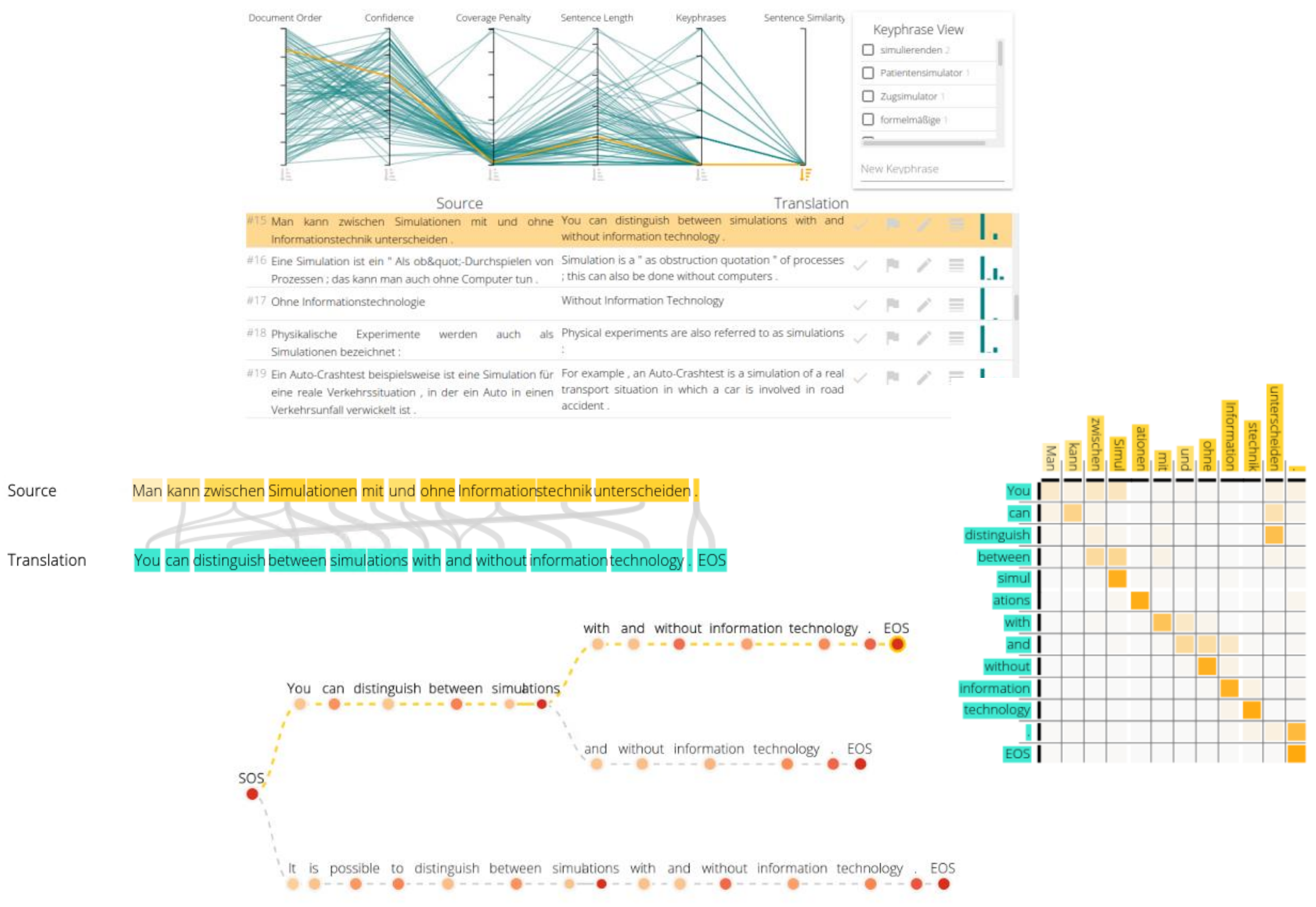
Positive Negative Uncertain Incorrect

Text classification for the example of imdb ratings

Input	Text sequences
Model	LSTMs
Strategy	Explore internal states
Visualization	Projection/distances of hidden states, expected prediction along the sequence
Interaction	Brushing and linking, tooltips with details
Features	Exploration, debugging

- **Neural Machine Translation (NMT)**

Our approach [2, 3] allows the translation of a document, provides information about the quality, and, for individual sentences, shows how the prediction for the translation was made (using a beam search visualization). The recommended prediction can be explored, adapted, and the quality of the model can also be improved by fine-tuning the model using the corrections of the user.



	Source	Translation
#15	Man kann zwischen Simulationen mit und ohne Informationstechnik unterscheiden .	You can distinguish between simulations with and without information technology .
#16	Eine Simulation ist ein " Als ob&quot;-Durchspielen von Prozessen ; das kann man auch ohne Computer tun .	Simulation is a " as obstruction quotation " of processes ; this can also be done without computers .
#17	Ohne Informationstechnologie	Without Information Technology
#18	Physikalische Experimente werden auch als Simulationen bezeichnet :	Physical experiments are also referred to as simulations :
#19	Ein Auto-Crashtest beispielsweise ist eine Simulation für eine reale Verkehrssituation , in der ein Auto in einen Verkehrsunfall verwickelt ist .	For example , an Auto-Crashtest is a simulation of a real transport situation in which a car is involved in road accident .

Source: Man kann zwischen Simulationen mit und ohne Informationstechnik unterscheiden .

Translation: You can distinguish between simulations with and without information technology . EOS

SOS
You can distinguish between simulations
with and without information technology . EOS
and without information technology . EOS
It is possible to distinguish between simulations with and without information technology . EOS
Man
kann
zwischen
Simul
ationen
mit
und
ohne
Information
stechnik
unterscheiden
.
You
can
distinguish
between
simul
ations
with
and
without
information
technology
.
EOS

Neural machine translation system

Input	Document with sentences
Model	LSTM/Transformer
Strategy	Explore prediction results and internal states
Visualization	Quality metrics, attention, beam search
Interaction	Find incorrect sentences, user corrections, brushing and linking
Features	Exploration, debugging, correct translations, improve model by fine-tuning

Supplemental Material

You can find a collection of videos and more screenshots of our approaches on DaRUS [4].

Future Work

- Apply our methods in an adapted form to other simulation-related data
- Develop new methods to explain AI models in other language-related areas and beyond
- Investigate Visual Question Answering (VQA), where we also intend to make internal states visible to users such that they better understand prediction results

References

[1] R. Garcia, T. Munz, D. Weiskopf. Visual analytics tool for the interpretation of hidden states in recurrent neural networks. VCIBA, 2021.

[2] T. Munz, D. Väth, P. Kuznecov, T. Vu, D. Weiskopf. Visual-interactive neural machine translation. Graphics Interface, 2021.

[3] T. Munz, D. Väth, P. Kuznecov, T. Vu, D. Weiskopf. Visualization-based improvement of neural machine translation. Computers & Graphics, 2022.

[4] T. Munz-Körner, S. Künzel, D. Weiskopf. Supplemental material for "Visual-explainable AI: The use case of language models". DaRUS, V1, 2023.

www.simtech.uni-stuttgart.de

Visualization Research Center
University of Stuttgart

SimTech