



University of  
Zurich<sup>UZH</sup>

École Polytechnique Fédérale de Lausanne

# Personalizing Automatic Text Simplification for Persons with Cognitive Impairments using Direct Preference Optimization

by Kaede Johnson

## Master Thesis

Supervisors:

Dr. Yingqiang Gao, University of Zurich

Dr. Matthieu Salzmann, EPFL

**Submitted to:**

EPFL  
College of Humanities  
CM 2 267 (Bâtiment CM)  
Station 10  
CH-1015 Lausanne

**Hosted at:**

University of Zurich  
Department of Computational Linguistics  
Language, Technology and Accessibility  
Andreasstrasse 15  
CH-8050 Zurich

May 11, 2025

Dedicated to Kirk Johnson,  
whose Bartokian scacchic unraveling following but a few casual games  
reminded me of what awesome power lies in resilience.

# Acknowledgments

I thank Dr. Yingqiang Gao of the University of Zurich (UZH) Department of Computational Linguistics for his generous and superlative supervision during this thesis. I thank Professor Sarah Ebling for the guidance and resources I was offered during my time in the UZH Language, Technology and Accessibility Group. I thank David Froehlich of capito Austria for his assistance in contacting and organizing target users, text simplification experts, and session proctors for preference pair annotation sessions. I thank Luisa Carrer of the Zurich University of Applied Sciences (ZHAW) for reviewing annotation session strategy and training documents, providing guidance on prompt creation, and recommending a text simplification expert. I thank Ursula Semlitsch and the other annotation session proctors at capito Austria for facilitating preference data acquisition. I thank Lisa Arter, Alessia Battisti, Sophia Conrad, Lukas Fischer, (again) Yingqiang Gao, Patrick Haller, Hanna Hubarava, Annette Rios, Anja Ryser, Patricia Scheurer, and Chiara Tschirner for manually reviewing simplifications and creating pairs for preference annotation. I thank Guglielmo Chelazzi Grandinetti and Igor Mustac for hosting our web application on UZH IT infrastructure and enabling external access for annotators. I thank Dr. Martin Kappus of ZHAW for providing feedback on our web application and annotation session strategy. I thank Léa Gainon for reviewing the French-language *Résumé*. I thank my peers in AND 2.25 for enriching my life throughout the duration of this project. Finally, I thank the target group annotators and text simplification experts for providing the preferences that made this research original.

*Oerlikon, Zurich, May 11, 2025*

Kaede Johnson

# Abstract

Automatic text simplification (ATS) aims to enhance language accessibility for several target groups, including persons with cognitive impairments. Recent advancements in generative AI, especially large language models (LLMs), have substantially improved the quality of ATS, thereby mitigating information barriers for target group persons. However, existing LLM-based ATS systems frequently overlook feedback from the target group during development and may fail to tailor output to user needs as a result. In this work, we extend the standard supervised fine-tuning (SFT) approach for adapting LLM-based ATS models by leveraging a computationally efficient LLM alignment technique—direct preference optimization (DPO). Specifically, we generate ATS pairs with mainstream LLMs, collect human feedback for these pairs from text simplification experts and persons with cognitive impairments, and use this feedback to post-train LLM-based ATS models using DPO. We find that DPO can successfully personalize ATS output provided preferences are consistent at the group level. Additionally, we find that preference consistency is more important than model- or content-related factors when personalizing ATS with DPO. This work represents a step towards personalizing inclusive AI systems and underscores the need to adapt research designs when involving target group participants in AI development.

**Keywords:** Automatic Text Simplification, Cognitive Impairments, Direct Preference Optimization, Personalization, Preference Alignment, Accessibility, HITL, LLM

# Résumé

La simplification automatique des textes (ATS) vise à améliorer l’accessibilité du langage pour plusieurs groupes cibles, notamment les personnes souffrant de déficiences cognitives. Les progrès récents de l’IA générative, en particulier les modèles de langage (LLM), ont considérablement amélioré la qualité de l’ATS, réduisant ainsi les obstacles à l’information pour les personnes du groupe cible. Cependant, les systèmes existants basés sur les LLM négligent souvent le retour du groupe cible pendant le développement et ne parviennent donc pas à adapter les résultats aux besoins de l’utilisateur. Dans ce travail, nous étendons l’approche standard de réglage fin supervisé (SFT) pour adapter les modèles ATS basés sur des LLM en tirant parti d’une technique d’alignement des LLM efficace sur le plan computationnel - l’optimisation directe des préférences (DPO). Plus particulièrement, nous générons des paires d’ATS avec des LLM génériques, recueillons des commentaires humains pour ces paires de la part d’experts en simplification de texte et de personnes souffrant de déficiences cognitives, et utilisons ces commentaires pour post-entraîner des modèles ATS basés sur des LLM à l’aide de la DPO. Nous constatons que la DPO peut personnaliser avec succès la sortie ATS à condition que les préférences soient cohérentes au niveau du groupe. En outre, nous trouvons que la cohérence des préférences est plus importante que les facteurs liés au modèle ou au contenu lors de la personnalisation de l’ATS à l’aide de la DPO. Ce travail représente une étape vers la personnalisation de systèmes d’IA inclusifs et souligne la nécessité d’adapter les modèles de recherche lors de l’implication de participants de groupes cibles dans le développement de l’IA.

# Contents

<b>Acknowledgments</b>	<b>1</b>
<b>Abstract (English/Français)</b>	<b>2</b>
<b>1 Introduction</b>	<b>6</b>
1.1 Background . . . . .	6
1.2 Motivation . . . . .	7
<b>2 Related Works</b>	<b>9</b>
2.1 German ATS Research . . . . .	9
2.2 LLM Personalization . . . . .	10
<b>3 Implementation</b>	<b>13</b>
3.1 Pipeline Overview . . . . .	13
3.2 Supervised Fine-Tuning (SFT) . . . . .	14
3.2.1 SFT Training . . . . .	14
3.2.2 SFT Evaluation . . . . .	21
3.3 Preference Pairs: HF4ATS-DPO . . . . .	26
3.3.1 ATS Pair Creation . . . . .	26
3.3.2 ATS Pair Annotation . . . . .	33
3.4 Direct Preference Optimization . . . . .	37
3.4.1 Dataset Construction . . . . .	37
3.4.2 DPO Training . . . . .	39
<b>4 Evaluation</b>	<b>40</b>
4.1 Datasets and Model Checkpoints . . . . .	40
4.2 Automatic Evaluation . . . . .	41
4.3 Human Evaluation . . . . .	42
<b>5 Results and Discussion</b>	<b>44</b>
5.1 Preference Pair Annotations . . . . .	44
5.2 DPO . . . . .	46
5.2.1 Quality Assessment of Generated ATS . . . . .	47

5.2.2	Impact of Individual Factors on DPO Post-training . . . . .	48
5.2.3	Personalization Success Rates of Target and Expert Models . . . . .	50
<b>6</b>	<b>Conclusion and Future Work</b>	<b>54</b>
6.1	Conclusion . . . . .	54
6.2	Limitations and Future Work . . . . .	54
	<b>Bibliography</b>	<b>56</b>
	<b>Appendices</b>	<b>63</b>
<b>A</b>	<b>SFT Grid Search for Parameter Mix</b>	<b>64</b>
<b>B</b>	<b>ATS Pair Creation Filters</b>	<b>66</b>
<b>C</b>	<b>Annotation Session Documents</b>	<b>71</b>
<b>D</b>	<b>Web Application</b>	<b>85</b>
<b>E</b>	<b>Problematic Pairs</b>	<b>87</b>
<b>F</b>	<b>Preference for Simple vs. Complex</b>	<b>103</b>

# Chapter 1

## Introduction

### 1.1 Background

The number of persons with cognitive impairments is estimated to be over 100 million worldwide [55, 64]. Although most of these impairments are classified as mild [55], they nonetheless erect barriers in key domains such as communication, information acquisition, and involvement in civic life. AI tools can be a means of reducing such barriers and enabling this demographic’s greater integration into society. Research topics aligned with this objective—or aimed more broadly at enhancing accessibility—include automatic audio description generation [19, 21], automatic sign language interpretation [44, 78], language sample analysis [53], and automatic text simplification [9, 31], the lattermost being the focus of this project.

Automatic text simplification (ATS) is a natural language processing (NLP) task that converts standard-language text into a more comprehensible form. As an open-ended task, it may add, delete, split, replace, or reorder content to improve text readability, increase lexical and syntactic simplicity, and optimize informational complexity [23, 71]. Target audiences for ATS include non-native language learners, persons with low literacy, and persons with cognitive impairments. This final group in particular encounters fundamental challenges in comprehending complex text; domain-specific jargon, implicit metaphors, advanced grammar or diction, and extraneous details restrict the processing of information flows relevant to daily life [57] and motivate the need for personalized ATS systems.

With recent advancements in large language models (LLMs), ATS systems have become more capable of generating accessible, high-quality text simplifications. The growing popularity of LLM-based ATS systems stems from their ease of implementation and performance supremacy relative to other off-the-shelf solutions. However, the perspectives of persons with cognitive impairments are frequently overlooked in the development of inclusive AI technologies [7], and they are seldom



consulted for their personal preferences on AI-generated output. The communication barriers associated with cognitive impairments also impose difficulties for target group persons seeking to express their own opinions [10], further complicating the gathering of accurate feedback. As a result, the involvement of persons with cognitive impairments is typically limited to a final evaluation phase, where they provide feedback on simplification systems that, in most cases, are not further refined.

We therefore identify two major gaps in ATS research:

**Research Gap 1** Persons with cognitive impairments are not constructively involved in the implementation phase of AI-driven ATS research.

**Research Gap 2** Despite being a distinct target group with specific needs, ATS models are rarely personalized for persons with cognitive impairments in particular.

Existing ATS training data is often curated by text simplification experts. Relying on this data overlooks the research gap related to the active participation of persons with intellectual impairments. Without specifying a target audience during curation, this data also overlooks the research gap related to audience targeting.

One candidate for addressing both research gaps is alignment. LLM alignment approaches such as reinforcement learning from human feedback (RLHF; [11]) chain LLM outputs to human values, expectations, and preferences [46]. Integrating RLHF into an LLM’s post-training pipeline could allow human feedback to guide text generation in a human-centered and inclusive direction. However, traditional RLHF methods such as Proximal Policy Optimization (PPO; [59]) are difficult to implement due to (1) the need for human preference data at scale; (2) the need to load three LLMs (reference model, reward model, and policy model) simultaneously during training; and (3) the difficulty in obtaining a sophisticated reward model from cumbersome and sensitive hyperparameter tuning (e.g. the clipping threshold for policy updates). Such limitations prevent the implementation of PPO in real-world applications, including the personalization of LLM-based ATS models for persons with cognitive impairments. A simpler and more lightweight alignment architecture is therefore preferred.

## 1.2 Motivation

In this work, we aim to personalize LLM-based ATS models for persons with cognitive impairments (henceforth referred to as the target group) using a lightweight, cost-effective, and human-in-the-loop (HITL; [43, 75]) technical framework. We investigate direct preference optimization (DPO; [51]), an LLM alignment algorithm that does not require explicit reward modeling, as our primary personalization methodology. We also aim to establish an ethical and effective personalization workflow for target group needs by adhering to the validate-annotate-evaluate HITL principle and

integrating target group participants throughout implementation phases.

The main contributions of our work are as follows: (1) We develop an inclusive, accessible, and user-friendly web application to collect human preference data from target group participants and text simplification experts, ensuring minimal cognitive demand during human-computer interaction; (2) We release HF4ATS, currently the largest dataset of German-language ATS preference pairs generated by mainstream LLMs and annotated by humans; (3) We release six LLM-based ATS models post-trained with DPO on HF4ATS from open source models; (4) We conduct experiments to examine the impact of both model selection and preference pair characteristics on DPO performance; (5) We perform an analysis of human evaluations on post-DPO inferences in the context of personalizing LLM-based ATS models.

The rest of this work is structured as follows: Section 2 reviews related literature on German ATS research and LLM personalization. Section 3 details all stages of our DPO implementation, including fine-tuning, pair creation, pair annotation, and post-training. Section 4 describes both automatic and human evaluation procedures. Section 5 presents the key research findings and provides an in-depth discussion of the experimental results. Finally, Section 6 summarizes the major takeaways of our study and highlights directions for future research.

## Chapter 2

# Related Works

### 2.1 German ATS Research

ATS research traditionally relied on rule-based [54, 63, 70] and statistical approaches [4, 50, 76] after its emergence as an NLP task in the late 1990s. Rule-based methods employ look-up tables for lexical and syntactic operations, while statistical approaches frame text simplification as a sequence-to-sequence task, often modeled using statistical machine translation. Recent advancements in LLMs have significantly transformed state-of-the-art ATS systems, enabling an end-to-end approach without feature engineering. This shift has progressed from encoder-decoder architectures to decoder-only models, driven by the enhanced computational efficiency and scalability of modern LLMs. As a result, decoder-only LLMs have become general-purpose problem solvers, redefining the learning paradigm for ATS.

While most ATS models were trained on English data [1, 58, 61], German ATS research has gained increasing attention in recent years, driven by active political and legal initiatives in German-speaking countries [15]. Notable examples include the *Barrierefreie-Informationstechnik-Verordnung* (*Accessible Information Technology Regulation*, BITV 2.0, 2011) in Germany, national action plans on disability in Austria [17], and the ratification of the *United Nations Convention on the Rights of Persons with Disabilities* (UN-CRPD) in Germany (2009), Austria (2008), and Switzerland (2014). These efforts have significantly advanced German ATS research, particularly in dataset construction [3, 6, 22, 33, 34, 56, 60, 68, 72], text alignment [66, 67], and model training [2, 25, 65].

Among existing German ATS models, BART [36], T5 [52], and their multilingual variants [38, 77] are the most commonly used base models, while benchmarking against commercial LLMs like GPT-3.5 or GPT-4 has also become standard practice. In this work, we fine-tune and post-train the mainstream open-source LLMs Llama [14, 73] and Mistral [30] to evaluate their ability to adapt to human preferences on automatic simplifications.

## 2.2 LLM Personalization

As summarized in the literature, there are three stages in the traditional route toward superlative text generation [48]:

**Large-scale Pre-training** In this stage, autoregressive LLMs (also known as causal LLMs) acquire knowledge from large-scale corpora through self-supervised next-token prediction. This process optimizes a cross-entropy objective to maximize the accumulated log-likelihood, enabling the pre-trained language models (PLMs) to develop deeper linguistic and contextual understanding of language.

**Supervised Fine-tuning (SFT)** PLMs undergo further fine-tuning on task-specific datasets to improve task comprehension and instruction adherence. The resulting fine-tuned models are optimized for both task execution and instruction following, enhancing their effectiveness in downstream NLP applications.

**Preference Learning** The final stage involves post-training fine-tuned models on human preference data to align their generations with human expectations. This facilitates responses that are as ethical, helpful, and informative as possible, and these dividends are largest when personalization is employed to target specific end users.

Preference learning, together with personalized prompting and personalized adaptation, constitutes one of the three key approaches to personalizing LLMs [37]. It is the only approach of the three that directly incorporates subjective human feedback into the learning process [81]. Furthermore, personalized prompting and personalized adaptation require explicit user profile data—e.g. gender, social relationships, occupation, and personal interests [37]—to guide text generation, data which cannot be obtained for persons with cognitive impairments due to ethical and legal constraints. Preference learning requires no such data, and the human feedback data it does rely on can be more easily anonymized.

Traditional Reinforcement Learning from Human Feedback (RLHF) methods such as Proximal Policy Optimization (PPO; [59]) utilize preference data to train an explicit reward model that guides the alignment process. As an on-policy RLHF approach which requires on-the-fly training data generation, PPO has demonstrated remarkable performance in conversational and coding tasks. However, it requires loading three LLMs simultaneously—a reference model  $\pi_{\text{ref}}$ , a policy model  $\pi_{\theta}$ , and a reward model  $R_{\psi}(\cdot)$ —making it particularly computationally intensive and difficult to implement. Some studies [45, 80] have explored the use of reinforcement learning within lightweight encoder-decoder generation pipelines for ATS. However, these still assume an explicit reward function that integrates supervised training signals without human preferences.

On the other hand, methods such as Direct Preference Optimization (DPO; [51]) eliminate the need for explicit reward modeling by learning directly from preference data. This makes them

significantly more lightweight and computationally efficient than PPO [29]. Moreover, as an offline method, DPO enables the pre-curation of preference data, thereby streamlining data collection in a manner than can ensure target group individuals' direct involvement. Care must be taken when accumulating preference data, however, as research shows high-quality preference pairs are the primary factor driving performance improvements in DPO post-training [29].

Consider an ATS preference dataset  $\mathcal{D}$  consisting of triplets  $(x, y_w, y_l)$ , where  $x$  is a complex text,  $y_w$  is the LLM-generated simplification of  $x$  preferred by humans, and  $y_l$  is the LLM-generated simplification of  $x$  dispreferred by humans. DPO aims to train a policy model  $\pi_\theta$  that, given input  $x$ , assigns a higher preference score to  $y_w$  than  $y_l$ . This state can be modeled as a probabilistic ranking with the Bradley-Terry model [8]:

$$P(y_w > y_l | x) = \frac{\exp(R_\psi(x, y_w))}{\exp(R_\psi(x, y_w)) + \exp(R_\psi(x, y_l))} = \sigma(R_\psi(x, y_w) - R_\psi(x, y_l)), \quad (2.1)$$

where  $\sigma(\cdot)$  is the sigmoid function. By applying a reparameterization to the standard reinforcement learning reward function optimization, one can remove  $R_\psi(x, y)$  from reward modeling and estimate the implicit reward  $\hat{r}(x, y)$  directly from training samples [51]. The standard DPO training objective then becomes

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right], \quad (2.2)$$

where the parameter  $\beta$  regulates the deviation of policy model  $\pi_\theta$  from reference model  $\pi_{\text{ref}}$ , ensuring the log-odd differences remain within a controlled range. This log-odd difference between the preferred and dispreferred text simplification is the so-called implicit reward margin:

$$\hat{r}(x, y_w, y_l) = \beta \left( \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right), \quad (2.3)$$

which can be directly estimated from the training triplets. Post-training with DPO involves initializing both  $\pi_{\text{ref}}$  and  $\pi_\theta$  to the SFT model  $\pi^{\text{SFT}}$ , freezing all  $\pi_{\text{ref}}$  model weights, and applying the gradient  $\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$  via backpropagation to the policy model  $\pi_\theta$  alone.

We employ preference learning and DPO specifically to pursue LLM personalization for ATS. Due to ethical considerations, we post-train on data from multiple target users, aiming to improve ATS for the target group collectively. We accept possible diversity in the type or severity of annotators' cognitive impairments in light of these ethical considerations. Our approach relies solely on LLM-generated ATS preference pairs annotated by target group persons and text simplification experts, ensuring group-level personalization without any user profiling.

We propose the following Research Questions (RQs):

**RQ1** Can DPO post-training with pairwise human preferences further improve the quality of ATS, as measured by automatic evaluation metrics?

**RQ2** To what extent do preference inconsistency, information equality, LLM backbone, and preference source influence the effectiveness of DPO post-training?

**RQ3** Can DPO post-training enable the successful group-level personalization of ATS models despite these uncertainties?

In the next section, we outline the full research pipeline we implemented to answer the research questions above. This includes (1) a summary of our pipeline, (2) a description of our SFT phase, (3) a description of our ATS pair creation process, (4) a description of our ATS pair annotation process, and (4) a description of our DPO post-training.

## Chapter 3

# Implementation

### 3.1 Pipeline Overview

Our preference alignment implementation pipeline, visible in Figure 3.1, broadly followed the pipeline proposed by Rafailov et al. We first conducted Supervised Fine-Tuning (SFT) for ATS on four language models. The data used for SFT was manually-aligned sentence-to-sentence(s) data sourced from Austrian Press Agency (APA) articles. Following a grid search for optimal hyperparameters and the subsequent SFT, we retained three SFT model checkpoints for inference and preference alignment. Next, a team of pair creators reviewed ATS inferences generated from APA sentence inputs and created 3,037 ATS pairs according to a uniform set of criteria. We recruited 15 persons with cognitive impairments and 4 text simplification experts to indicate a preference within each pair, providing the preference data used for Direct Preference Optimization (DPO) on our three winning SFT checkpoints. Finally, we conducted DPO post training on a slate of preference pair subsets to determine if DPO could improve simplification quality and achieve personalization.

We call the dataset used in this project Human Feedback for Automatic Text Simplification (HF4ATS). It is composed of HF4ATS-SFT ( $\mathcal{D}_{\text{SFT}}$ ), a complex-simple sentence-to-sentence(s) dataset suitable for fine-tuning German LLM-based ATS models, and HF4ATS-DPO ( $\mathcal{D}_{\text{DPO}}$ ), an ATS preference pair dataset annotated by native German speakers both with and without cognitive impairments. To the best of our knowledge, HF4ATS-DPO is the first German-language preference dataset collected directly from the target group for this purpose.

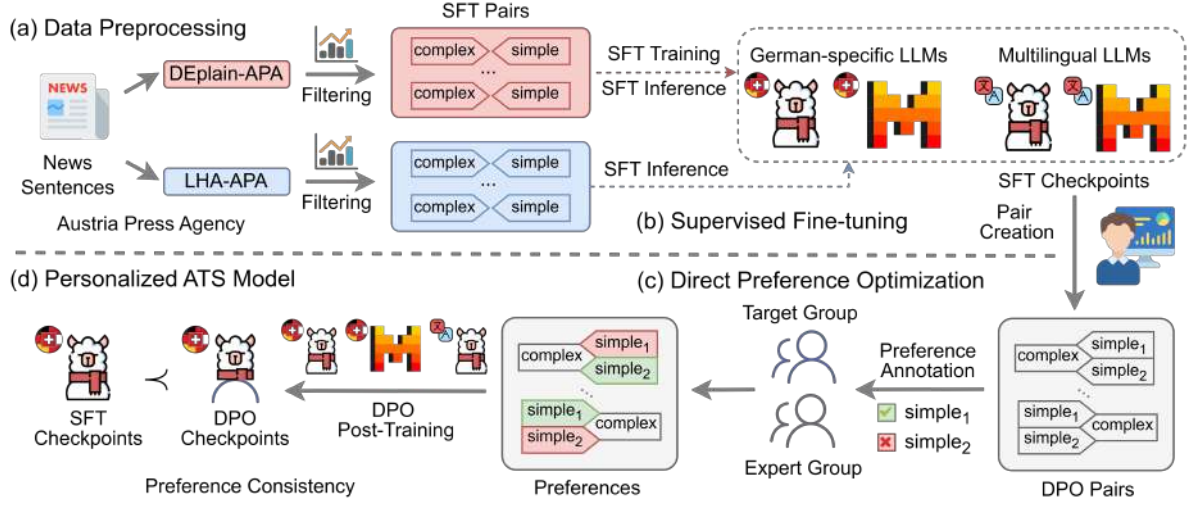


Figure 3.1: **Our personalization pipeline for LLM-based ATS models.** (a) **Data Filtering:** We selected high-quality sentence-level complex-simple pairs from two datasets; (b) **Supervised Fine-tuning:** We finetuned pre-trained German-specific and multilingual LLMs; (c) **Direct Preference Optimization:** We post-trained SFT checkpoints with human preference data collected from both target and expert group annotators; (d) **Evaluation:** We evaluated the DPO checkpoints against their SFT precursors for their alignment with human preferences.

## 3.2 Supervised Fine-Tuning (SFT)

The first step in our preference alignment pipeline was SFT. In the context of text simplification, SFT uses a preexisting dataset composed complex sentences and their simple counterparts to update model weights in pursuit of improved simplification quality. We implemented SFT because we sought to follow the alignment regime proposed by Rafailov et al., to format inferences in a manner conducive to preference pair annotation, and to expose our pre-trained models to the full diversity of text simplification approaches.

### 3.2.1 SFT Training

**Pretrained Model Selection** When selecting pretrained models for SFT and eventual preference alignment, we prioritized multilinguality, accessibility, popularity in contemporary NLP research, and prior instruction tuning. Our first criterion targeted models with German-language training data in particular given that our preference pair annotators would be German readers. Accessibility and popularity ensured our work would be reproducible while having the capacity to reflect the dividends of preference alignment. The instruction-tuned condition, meanwhile, would allow us to forgo the general instruction tuning step our models might otherwise need to generate suitable inferences.



We chose DiscoLeo-Llama-3-8B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct, and LeoLM-Mistral-7B-Chat in this work based on a balance of the above criteria. While both DiscoLeo-Llama-3-8B-Instruct and LeoLM-Mistral-7B-Chat may not be as popular as our other two models, we made an exception due to their larger share of German-language pretraining data. Regarding instruction tuning, LeoLM-Mistral-7B-Chat is the only model not described as an Instruct variant, but it was nonetheless trained on German instruction datasets and there are no alternative Mistral-7B-Instruct variants with a similar amount of German-language pretraining data. The share of preference pairs created with LeoLM-Mistral-7B-Chat inferences in a model-blind environment (see Section 3.3.1) confirm that it was capable of producing suitable inferences despite this pretraining difference.

**Raw SFT Data** HF4ATS-SFT data was sampled from DEPLAIN-APA [68], a collection of manually aligned complex-simple sentence-level pairs. The sentences in DEPLAIN-APA are sourced from Austria Press Agency (APA) news items written in Austrian German by text simplification professionals and published between May 2019 and April 2021. The dataset includes 483 document pairs, from which 13,122 sentences pairs were manually aligned by native German speakers. Common topics include political events, coronavirus, economic indicators, weather, sports, and crime. All DEPLAIN-APA pairs align sentences from articles classified as B1 under the Common European Framework of Reference for Languages (CEFR) with sentences from articles classified as A2.

A sample data point is as follows:

- Complex (from a B1 news article): *‘Am Mittwoch hat der Iran bekanntgegeben, dass er teilweise aus dem Atom-Abkommen aussteigt.’* (On Wednesday, Iran announced that it will partially withdraw from the nuclear agreement.)
- Simple (from an A2 news article): *‘Der Iran wird teilweise aus dem Atom-Abkommen aussteigen.’* (Iran will partially withdraw from the nuclear agreement.)

**Data Filtering** Before using DEPLAIN-APA to conduct SFT, we filtered the 13,122 sentence pairs based on the number of complex sentences simplified and the level of similarity between the complex and simple sentence(s).

Part of our filtering motivation was related to sentence alignment. In general, a single complex sentence may be simplified into just one simple sentence (one-to-one alignment) or multiple simple sentences (one-to-many alignment). It is also possible to simplify multiple complex sentences at once (many-to-one or many-to-many alignment). However, just 6.4% of DEPLAIN-APA pairs constitute many-to-one or many-to-many alignment. Rather than rely on so few complex-simple pairs to incorporate many-to-many simplification in our research, we removed these 850 data points and focused exclusively on one-to-many or one-to-one sentence simplification.

Unfortunately, with a focus narrowed to the sentence level, many alignments in DEPLAIN - APA forgo important contextual information. This means complex sentences in DEPLAIN - APA do not necessarily entail their simple counterparts, as seen in the following data point:

- Complex: *‘Es gibt aber große Unterschiede.’* (But there are big differences.)
- Simple: *‘Nicht in jedem Vanille-Eis ist gleich viel Luft drin.’* (Not every vanilla ice cream contains the same amount of air.)

In this example, the information contained in the simple sentence relies on context from the source article that is not contained in its paired complex sentence. This happens because pairs in DEPLAIN - APA, despite being manually written and aligned, are composed of sentences from disparate articles rather than a strict sentence-level simplification task. This likely complicated alignment for those who created the sentence pairs. The danger of including too many such data points during SFT is that models may learn to hallucinate context and violate entailment. At the same time, adding information is a valid approach to text simplification, and we therefore did not wish to remove the possibility entirely.

To curtail the most egregious non-entailment, we filtered out 614 pairs with a complex-simple embedding cosine similarity less than 0.5. We created vector representations for DEPLAIN - APA sentences using a German-tuned Sentence-BERT model [42] and calculated cosine similarity between the complex sentence embedding and the embedding for the simple sentence(s) within each sentence pair. The distribution of cosine similarities is shown in Figure 3.2 along with our lower bound of 0.5; our choice of 0.5 was heuristic and motivated in the final paragraph of this section.

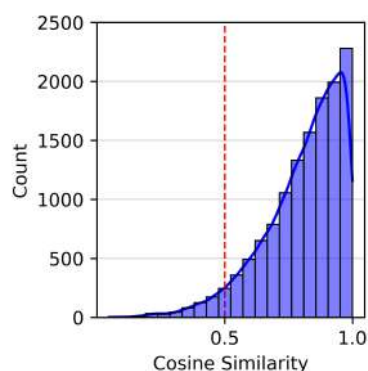


Figure 3.2: Distribution of cosine similarity between complex and simple sentences in DEPLAIN pairs. We apply a lower bound threshold at 0.50 in order to excise pairs that lack entailment.

At the other end of the spectrum, some simple sentences in DEPLAIN - APA nearly equate their complex counterpart, as seen in the following example:

- Complex: *‘Integration bedeutet, also dass jemand dazugehört.’* (Integration means that someone belongs.)

- Simple: *‘Integration bedeutet also, dass jemand dazugehört.’* (Integration means that someone belongs.)

The danger of including such intra-pair similarity during SFT is that models may learn to change very little about complex sentences. A model which learns to change little about a complex sentence may not exhibit enough diversity in task completion to permit informative preference pairs. Additionally, because DEPLAIN-APA data is later used for evaluation, we did not seek to include sentences which are not, according to DEPLAIN-APA, in need of simplification.

We removed 2,360 data points with ROUGE-1, -2, and -L F1 scores between complex sentence (ROUGE reference) and simple sentence(s) (ROUGE output) higher than 0.8. The distribution of ROUGE scores is shown in Figure 3.3.

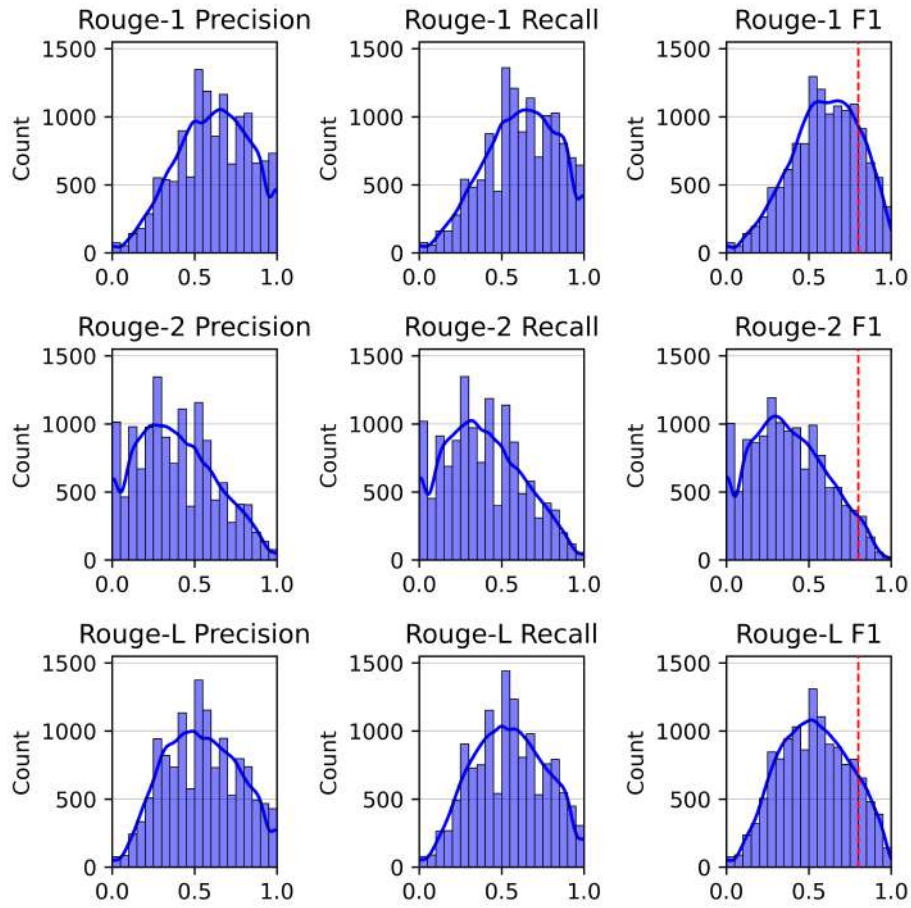


Figure 3.3: Distributions for ROUGE scores between complex and simple sentences in DEPLAIN. We apply an upper bound to F1 scores at 0.80.

The thresholds on embedding cosine similarity and ROUGE F1 scores were not intended to

be robust in the SFT preprocessing stage. As described in Section 3.3.1, we relied on humans to detect non-entailment and evaluate simplification quality during ATS pair creation. It was therefore sensible to apply relaxed, heuristic filters in the SFT preprocessing stage so our human pair creators could have access to a diverse pool of automatic simplifications during pair creation.

**Train, Development, and Test Sampling** We sampled 5,200 of the 9,414 remaining DEPLAIN-APA pairs and randomly assigned 3,600 to an SFT training subset, 800 to an SFT development subset, and 800 to a test subset for later evaluation. These 5,200 pairs constitute HF4ATS-SFT. The global subset of 5,200 data points was sampled using Gaussian weighting based on the length of the complex sentence, centered on 15 words (the average complex sentence word count in DEPLAIN is 11.24). Specifically, for a given complex sentence  $x$ , the sampling weight  $w_x$  was defined as

$$w_x = \exp\left(-\frac{(|x| - 15)^2}{2 \cdot \sigma^2}\right), \quad (3.1)$$

where  $|x|$  denotes the word count of the complex text and the standard deviation  $\sigma$  was set to 3.

We inflated the average length of complex sentences during sampling to boost the diversity of inferences shown to pair creators and, by extension, pair annotators. To arrive at this mechanism, we reason that, all else equal, and while respecting entailment, there are more ways to simplify a longer sentence than a shorter one. This is because less information may be removed from the latter. That said, to avoid encouraging long inferences that might induce high cognitive load, we only permitted data points where the simplification was less than 30 words in length.

The result, shown in Figure 3.4, was a sample with a word count distribution less skewed and higher on average for both complex sentences (left) and simple sentences (right).

From the remaining DEPLAIN-APA data points we sampled a further 2,580 pairs at random for use in one- and two-shot prompts in the train, development, and test subsets of HF4ATS-SFT. We then combined the remaining non-sampled pairs with data previously filtered out due to low embedding cosine similarity or high ROUGE F1 scores and preserved this set of 5,438 DEPLAIN data points for use in DPO inference.

**Prompting** We used a set of ten prompts during SFT training. Eight of the ten prompts were again used during ATS pair creation, DPO post-training, and final evaluation.

Our prompts, developed in collaboration with an automatic text simplification expert, approached the simplification task from a variety of angles. Each item below appeared in at least one prompt variant:

- Description of the target audience (Austrian persons with cognitive impairments).

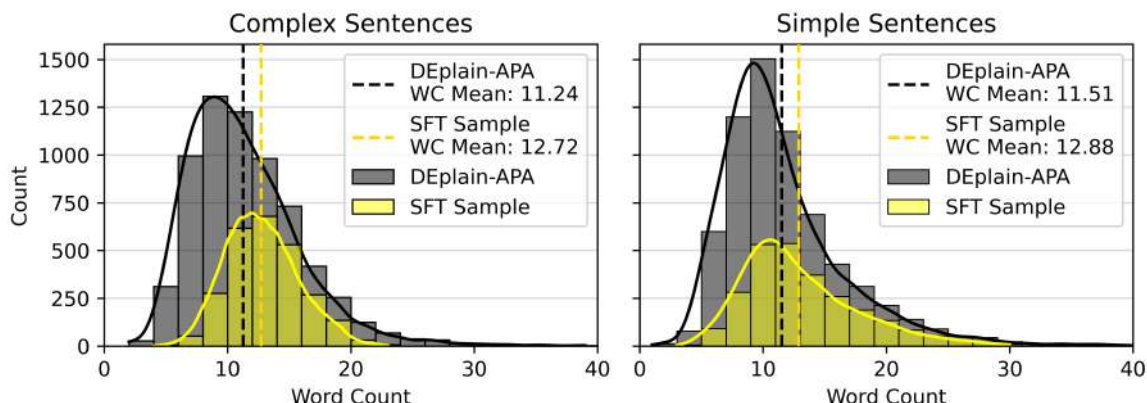


Figure 3.4: Word count distribution of complex sentences (left) and simple sentences (right) for data in DEPLAIN-APA and the 5,200-data-point sample drawn from DEPLAIN-APA for SFT train, development, and test. For complex sentences, the pre-sample distribution exhibits more skew and a lower mean word count than the weighted sample.

- Target goal of ‘*Leichte Sprache*’ (Simple Language).
- One-Shot prompting.
- Two-Shot prompting.
- Text simplification approaches (adding, removing, reordering, replacing, and splitting information) as sourced from the *Deutsches Institut für Normung’s* “*Empfehlungen für Deutsche Leichte Sprache*” (Recommendations for German Easy Language)<sup>1</sup> [13]

We favored zero-shot prompting to avoid biasing our model toward specific methods for carrying out the open-ended task of text simplification. That said, the lattermost two prompts below—which were only used in SFT training—provided in-context examples sampled without replacement from DEPLAIN-APA as described above. All prompts included the sentence ‘*Bitte gib nur eine Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare,*’ (Please provide only a simplification, without an introduction, alternatives, or comments) for formatting purposes. Prompts were assigned at random in all cases.

The complete set of prompts was as follows:

1. Schreibe den folgenden Satz in Leichter Sprache um: <complex\_sentence>. Bitte gib nur eine Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare.

<sup>1</sup><https://www.din.de/de/mitwirken/normenausschuesse/naerg/e-din-spec-33429-2023-04-empfehlungen-fuer-deutsche-leichte-sprache--901210>

2. Vereinfache den folgenden Satz, sodass Menschen mit kognitiver Beeinträchtigung den vereinfachten Satz verstehen können: <complex\_sentence>. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare.
3. Schreibe den folgenden komplexen Satz um und verwende einfachere Wörter, kürzere Sätze und reduzierte grammatikalische Strukturen. Der Inhalt und die Bedeutung sollen nach dem Umschreiben unverändert bleiben. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Komplex: <complex\_sentence>. Leicht:
4. Formulieren Sie den komplexen Satz um, indem Sie mindestens einen neuen einfachen Satz bilden. Behalten Sie die gleiche Bedeutung des Ausgangssatzes bei. Geben Sie bitte nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Komplex: <complex\_sentence>. Leicht:
5. Schreibe den folgenden komplexen Satz in Leichter Sprache um. Die Vereinfachung soll kurz und von geringer Komplexität sein (durchschnittlich acht bis fünfzehn Wörter pro Satz) und eine geringe Anzahl von Aussagen pro Satz enthalten. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Komplex: <complex\_sentence>. Leicht:
6. Schreibe den folgenden komplexen Satz in Leichter Sprache um. Die Wörter in deiner Vereinfachung sollen kurz, beschreibend, und häufig verwendet von Menschen mit kognitiver Beeinträchtigung sein. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Komplex: <complex\_sentence>. Leicht:
7. Schreibe den folgenden komplexen Satz in Leichter Sprache um. Deine Vereinfachung soll für Menschen mit kognitiver Beeinträchtigung in Österreich verständlich sein. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Komplex: <complex\_sentence>. Leicht:
8. Schreiben Sie den folgenden komplexen Satz in Leichter Sprache um. Sie können 1) den Satz in mehrere Sätze aufteilen, 2) die Wortstellung ändern, um die Grammatik zu vereinfachen, 3) Wörter hinzufügen, um schwierige Konzepte zu erklären, 4) Wörter, die sich mit unnötigen Informationen zusammenhängen, entfernen, und 5) schwierige Wörter durch einfachere Vokabeln ersetzen. Achten Sie darauf, dass der Satz leichter verständlich bleibt, ohne die Bedeutung zu verändern. Bitte geben Sie nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Komplex: <complex\_sentence>. Leicht:
9. Schreibe den folgenden komplexen Satz in Leichter Sprache um. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Hier ist ein Beispiel. Komplex: <complex\_sentence1>. Leicht: <simple\_sentence1>. Schreibe deine Vereinfachung nach "Leicht:". Komplex: <complex\_sentence>. Leicht:
10. Schreibe den folgenden komplexen Satz in Leichter Sprache um. Bitte gib nur die Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare. Hier sind zwei Beispiele. Komplex: <complex\_sentence1>. Leicht: <simple\_sentence1>. Komplex: <complex\_sentence2>. Leicht: <simple\_sentence2>. Schreibe deine Vereinfachung nach "Leicht:". Komplex: <complex\_sentence>. Leicht:

Above, <complex\_sentence> refers to the sentence the language model should simplify. Meanwhile, <complex\_sentence1> and <complex\_sentence2> refer to complex sentences simplified

in the prompt, for which `<simple_sentence1>` and `<simple_sentence2>` serve as simple counterparts.

**Training Setup** We conducted SFT with the four language models, the 3,600 HF4ATS-SFT complex-simple train pairs, and the prompts described above. We performed evaluation every 400 data points (during cross-model comparison) or 448 data points (during our grid search). Our motivation for investigating intermediary checkpoints stemmed from Zhou et al., who find that optimal SFT performance may follow from only a couple thousand training data points [82].

To implement SFT, we used the TRL SFTTrainer library [74]. We left the pad token unchanged for DiscoLeo-Llama-3-8B-Instruct, set the pad token to `'<unk>'` for the two Mistral models, and set it to `'<|finetune_right_pad_id|>'` for Llama-3.1-8B-Instruct. The padding side was 'right' during all trainings. Our training regime incorporated prompt tokens into training loss (i.e., full-prompt tuning) as opposed to completion tokens only (completion-only tuning) in most cases. We prioritized this approach because (1) the average ratio of prompt length to completion length in our training data was above five and (2) our trainings incorporated a few thousand data points at a maximum rather than tens of thousands of data points. According to Shi et al., including prompt tokens in training loss calculation is either advantageous or at least neutral for open-ended tasks under these circumstances [62]. Nonetheless, for the sake of robustness, we included variants for Llama-3.1-8B-Instruct and DiscoLeo-Llama-3-8B-Instruct in which prompt tokens were not included in training loss calculation.

The stable components of our SFT setup included a paged AdamW optimizer [39] with weight decay of 0.01, a cosine annealing learning rate scheduler [40], max grad norm clipping at 1, a maximum sequence length of 300 tokens, a batch size of 16, and FP16 mixed precision. We performed parameter-efficient fine-tuning (PEFT; [24]) on a single A100 GPU using LoRA [26] with rank 16, a scaling factor of 32, and a dropout rate of 0.05. Our grid search for hyperparameters scanned across gradient accumulation step size (1, 2, and 4) and learning rate ( $1e-5$ ,  $5e-5$ , and  $1e-4$ ). We selected DiscoLeo-Llama-3-8B-Instruct for hyperparameter optimization due to its abundance of German-language training data relative to Llama-3.1-8B-Instruct and Mistral-7B-Instruct as well as its extensive instruction tuning compared to LeoLM-Mistral-7B-Chat.

### 3.2.2 SFT Evaluation

We needed to select from nine possible parameter combinations and 36 subsequent model checkpoints when deciding which checkpoints to carry forward to preference optimization. All evaluations in this section were conducted using the 800-data-point HF4ATS-SFT development sample described above.



**Automated Metrics** To inform our parameter mix and SFT checkpoint decisions, we calculated the following set of quality, readability, and quality control metrics for 800 inferences (we also used some of these metrics during preference optimization evaluation; see Section 4.2):

### 1. Simplification Quality

- (a) BERTScore [79]. This simplification metric calculates the cosine similarity of generated simplifications and their complex counterparts using BERT embeddings. We used the F1 score in our evaluation and HuggingFace’s Python library `evaluate` for implementation [74].
- (b) BLEU Score [47]. We included this n-gram-based simplification metric for completeness but de-prioritized it during our post-SFT model usage decision due to previous work finding it insufficient for text simplification evaluation [69]. We use HuggingFace’s Python library `evaluate` for implementation [74].
- (c) SARI [76]. This n-gram-based simplification metric aggregates metrics for edits performed in a generated simplification relative to a complex sentence and set of reference simplifications. We use HuggingFace’s Python library `evaluate` for implementation [74]. Note that there was just one reference simplification for all complex-simple data points in HF4ATS-SFT because the set of complex sentences in our DEPLAIN-APA subset was unique.
- (d) LENS [41]. This model-based simplification metric uses a RoBERTa embedding model trained on human numeric evaluations of sentence simplifications to predict simplification quality. Higher values indicate better simplifications. Note that the training data for LENS is based on English-language simplifications alone. Accordingly, we de-prioritized this metric during our post-SFT model usage decision. We used the LENS checkpoint and package listed on HuggingFace for implementation<sup>2</sup>.

### 2. Simplification Readability

- (a) Average Word Count. As elsewhere in this thesis, “word count” refers to the number of spaces in a sentence plus one.
- (b) Flesch Reading Ease [32]. This readability metric is a weighted average of average words per sentence and average syllables per word, subtracted from a constant. The scores range between 0 and 100, with higher values indicating greater readability. We calculated the metric for each generated simplification individually before averaging.
- (c) Wiener Sachtextformel (*Vienna Formula*), Variant 4 [5]. WSTF<sub>4</sub> is a reference-free readability metric for German texts defined as follows:

$$\text{WSTF}_4 = 0.2744 \times MS + 0.2656 \times SL - 1.693, \quad (3.2)$$

---

<sup>2</sup>Available at <https://huggingface.co/davidheineman/lens>.



where  $MS$  represents the percentage of words containing more than three syllables and  $SL$  denotes the average word count per sentence. A  $WSTF_4$  score of 4 indicates a very simple text, while a score of 15 indicates a very complex text. We selected  $WSTF_4$  as the most salient readability metric over Flesch Reading Ease because it was specifically designed for non-fiction, German-language text.

Both Flesch Reading Ease and  $WSTF_4$  were computed using the Python package `Textstat`<sup>3</sup>, configured for German language settings.

### 3. Training Regime Quality Control

- (a) Proportion of simplifications equal to their complex counterpart. To test for equality, we compared the lowercase version of two sentences after removing all non-alphabetic characters.
- (b) Proportion of empty generations. We considered any simplification less than 10 characters in length to be “empty”.
- (c) Proportion of simplifications classified as German-language. We used the highest-probability language reported by the Python package `langdetect`<sup>4</sup> to classify a simplification’s language. We included this metric due to the usage of multilingual language models in our research.
- (d) Cross-Entropy Loss.

We used greedy decoding in all cases. For those metrics which required reference sentences, we used the complex sentence’s corresponding simple sentence(s) in the HF4ATS-SFT development subset. The automatic metrics we prioritized were SARI for simplification quality and Wiener Sachtextformel for readability. We prioritized these two metrics because researchers created the former specifically for text simplification and the latter for non-fiction German-language text.

**Grid Search Results** We performed evaluation with the HF4ATS-SFT development subset every 448 training data points during our grid search. The results are shown in Appendix Figure A.1.

The parameter mix with learning rate  $1e-4$  and gradient accumulation step size 1—indicated by the solid line with circular markers in Appendix Figure A.1—attained the maximal SARI score. During the later stages of training, the same parameter mix settled at the lowest Wiener Sachtextformel grade level in the set. We therefore selected this parameter mix for cross-model comparison.

Other metrics generally supported this decision. Our optimal parameter mix exhibited relatively high Flesch Reading Ease, a relatively low proportion of inferences equal to the complex sentence, no empty inferences, and an adequate average BERTScore between complex and simple sentences.

---

<sup>3</sup>Available at <https://github.com/textstat/textstat>, MIT license.

<sup>4</sup>Available at <https://pypi.org/project/langdetect/>, Apache Software License.

Meanwhile, there was no clear parameter mix “winner” regarding average inference word count and language classification, and our optimal parameter mix’s low BLEU values were not concerning due to the aforementioned de-prioritization of BLEU for our evaluation.

**SFT Checkpoint Comparison and Selection** Following the results of our grid search, we trained all four models using a gradient accumulation step size of 1 and a learning rate of  $1e-4$ . Figure 3.5 displays the results, including those for the completion-only loss training variants using Llama-3.1-8B-Instruct or DiscoLeo-Llama-3-8B-Instruct.

No model exhibited a uniform advantage over the others. Considering only the four trainings with full-prompt loss tuning, the tuned Llama models reported high SARI while the Mistral models reported a frequent if small advantage in readability. The Llama models were also 5-6 times more likely than the Mistral models to produce output effectively equal to the complex sentence, indicating they were less likely to attempt the task.

The two models with completion-only loss tuning exhibited a marked relative increase in the average number of words per simplification. This is especially true for DiscoLeo-Llama-3-8B-Instruct. This boost in word count artificially inflated the two models’ advantage in quality and readability scores relative to the four trainings previously discussed. That said, adding or splitting information is a viable simplification strategy, so we did not dismiss models with a propensity for longer simplifications outright.

We retained three separate model checkpoints for preference optimization due to the inconsistency in relative performance across our six trainings. First, in pursuit of SARI maximization, we retained the full-prompt-loss DiscoLeo-Llama-3-8B-Instruct checkpoint at 2,800 training observations. Next, to represent the training framework with completion-only loss tuning, we retained the completion-only Llama-3.1-8B-Instruct checkpoint at 2,400 training observations. We retained this particular checkpoint because the high-variance evaluation metrics showed a local peak in BERTScore, a local dip in average simplification length, and a local dip in the proportion of simplifications equal to the complex sentence at 2,400 observations, all at no cost to SARI. Finally, to emphasize readability and round out our set with a non-Llama-based model, we retained the LeoLM-Mistral-7B-Chat checkpoint after 1,600 training observations. The two Mistral models perform similarly; we opted for the German-tuned variant due to slightly better readability according to Wiener Sachtextformel. Our motivation for selecting the checkpoint at 1,600 training observations lies not only in minimizing Wiener Sachtextformel grade level but also in protecting against potential hallucination during inference. Recall from Section 3.2.1 that our SFT data often exhibits lack of entailment due to absent context. Assuming our preprocessing filters did not completely excise this issue, we reasoned that retaining a model checkpoint trained with fewer data points (and therefore fewer entailment-violating sentence pairs) could serve as an inference source less likely to hallucinate.

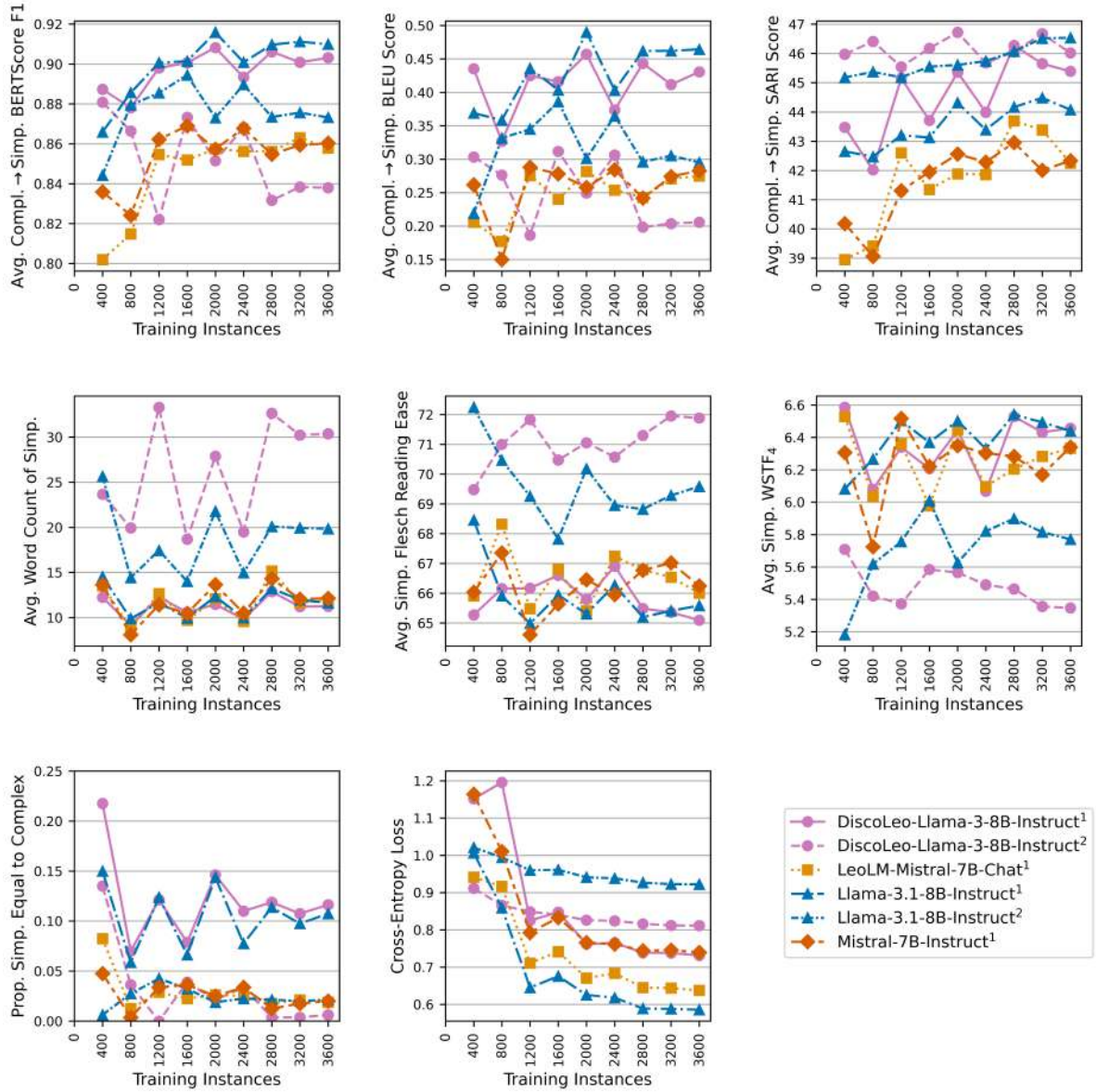


Figure 3.5: **Cross-model comparison for SFT checkpoint evaluation.** This figure compares different SFT configurations across multiple models using two loss computation strategies. **Full-prompt loss** (denoted as 1 in the legend) includes both instruction and completion tokens in the loss calculation, while **completion-only loss** (denoted as 2) considers only the completion tokens. This distinction reflects different assumptions about how supervision signal should be distributed over the input. We selected three different checkpoints for DPO post-training based on this evaluation: DiscoLeo-Llama-3-8B-Instruct (full-prompt loss) after 2,800 training observations, Llama-3.1-8B-Instruct (completion-only loss) after 2,400 training observations, and LeoLM-Mistral-7B-Chat after 1,600 training observations.

### 3.3 Preference Pairs: HF4ATS-DPO

We introduce HF4ATS-DPO: a twice-annotated set of approximately 3,000 German-language ATS preference pairs. Each preference pair contains a unique original (i.e. complex) sentence and two distinct automatic simplifications for that original sentence. We trained a team German speakers to create the ATS pairs using inferences from our three SFT model checkpoints. We then recruited a group of Austrian persons with cognitive impairments and text simplification experts to separately annotate the full set of preference pairs. The annotated preference pairs allowed us to implement and evaluate preference alignment on our three models using two distinct annotation sets.

#### 3.3.1 ATS Pair Creation

**Inference Raw Data** We drew complex sentences from two APA data sources when conducting inference for pair creation. First, we accessed complex sentences from the 5,438 `DEPLAIN-APA` data points not sampled into HF4ATS-SFT during SFT training, which we denote  $\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}$ . Second, we accessed sentences from original, non-simplified APA articles provided by APA-LHA [65], denoted  $\mathcal{D}_{\text{LHA}}$ .

APA-LHA is a sentence-level alignment dataset containing APA news items that cover the same topics as `DEPLAIN-APA`: political events, economic indicators, sports, etc. In fact, some `DEPLAIN-APA` sentences also appear in APA-LHA. However, unlike `DEPLAIN-APA`, APA-LHA contains sentences from non-simplified, original news items. This means it contains sentences classified above the B1 and A2 CEFR levels. We did not use APA-LHA during SFT because its underlying automatic alignment mechanism resulted in misaligned complex-simple pairs and its alignments do not distinguish between one-to-one, one-to-many, and many-to-one simplifications [68]. These issues would have exacerbated the entailment issues already present in `DEPLAIN-APA`'s manually aligned data. APA-LHA sentences were nonetheless suitable for preference pair inference because their topic distribution matched that of our SFT data. This topic similarity is beneficial for the original DPO pipeline [51], which prevents preference alignment from causing large deviations from the data distribution established during SFT. We also sought to include APA-LHA sentences during inference because we reasoned their relatively high complexity would enable our models to make more extreme changes during automatic simplification.

**Pre-Inference Filtering and Sampling** To create a set of complex sentences for inference, we began with the 11,078 APA-LHA sentences from non-simplified news items. We then removed unnecessary spaces, sentences that appeared in `DEPLAIN-APA`, and one English-language sentence before combining the subset with leftover `DEPLAIN-APA` complex sentences. From this global set we removed data points with sentence-ending punctuation anywhere except at the end of the sentence (in APA-LHA data, we found this occasionally indicated multiple sentences rather than a

single complex sentence). Finally, we sampled from the APA-LHA sentences and DEPLAIN - APA sentences to arrive at 8,000 complex sentences for use in ATS pair creation.

We applied strict length thresholds and weighted based on word count during our sampling. Our motivation was, once again, to match the SFT data distribution during preference pair creation. We first removed all sentences less than five words and greater than thirty words in length. We then applied Gaussian sampling with different weighting schemes to the two HF4ATS-DPO inference sources. From the DEPLAIN - APA subset we sampled a complex text  $x$  with a Gaussian weight  $w_x$  defined as

$$w_{x \sim \mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}} = \exp \left( - \frac{(|x| - \mu_{\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}})^2}{2 \cdot \sigma^2} \right), \quad (3.3)$$

where the mean

$$\mu_{\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}} = \frac{\sum_{x' \in \mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}} |x'|}{|\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}|} \quad (3.4)$$

corresponds to the average word count of complex texts from the leftover DEPLAIN subset, with  $|\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}|$  denoting the subset's size and  $|x|$  denoting a given complex text's word count. From APA-LHA,  $x$  was sampled with weight  $w_x$  defined as

$$w_{x \sim \mathcal{D}_{\text{LHA}}} = \exp \left( - \frac{(|x| - (\mu_{\mathcal{D}_{\text{LHA}}} + \eta \cdot (\mu_{\mathcal{D}_{\text{LHA}}} - \mu_{\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}}))^2}{2 \cdot \sigma^2} \right), \quad (3.5)$$

where the mean

$$\mu_{\mathcal{D}_{\text{LHA}}} = \frac{\sum_{x' \in \mathcal{D}_{\text{LHA}}} |x'|}{|\mathcal{D}_{\text{LHA}}|} \quad (3.6)$$

represents the average word count of complex texts from APA-LHA and  $\eta = 4,800/8,000$  is a scaling factor reflecting the share of LHA-APA (as opposed to leftover DEPLAIN) complex sentences present in the 8,000 data point inference set.

Figure 3.6 shows that these sampling weight specifications allowed the relatively short sentences from  $\mathcal{D}_{\text{DEPLAIN}} \setminus \mathcal{D}_{\text{SFT}}$  and relatively long  $\mathcal{D}_{\text{LHA}}$  sentences to, when combined, emulate the training data from HF4ATS-SFT.

**Inference Parameters** We generated 20 inferences for all 8,000 complex sentences using four top-p sampling decoding parameter sets:

1. Temperature = 1, p = 0.9
2. Temperature = 1.3, p = 0.8

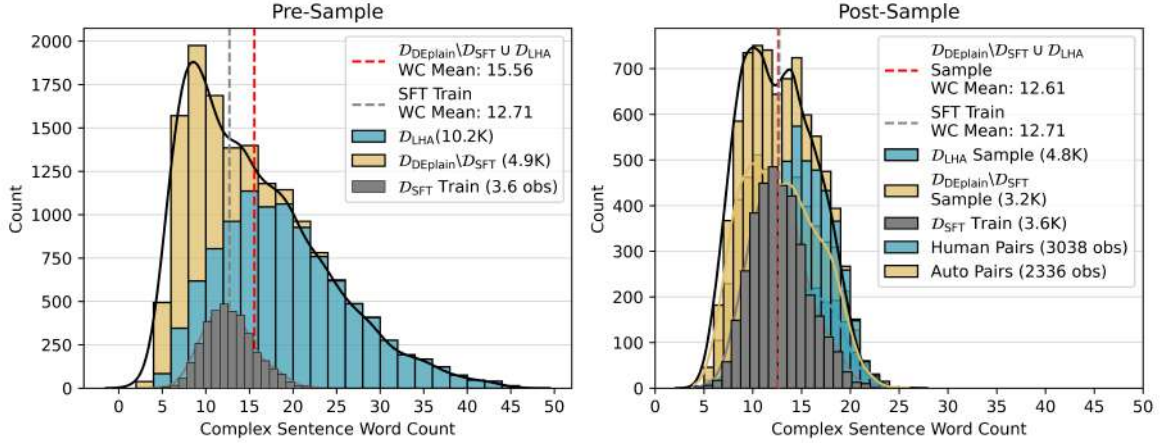


Figure 3.6:  $D_{\text{EPLAIN}} - \text{APA}$  and  $\text{APA} - \text{LHA}$  complex sentence word count distributions, pre-sample (left) and post-sample (right). The weighted sample centers the word count distribution around the  $D_{\text{EPLAIN}}$  SFT train dataset, an optimal condition for preference alignment.

3. Temperature = 1.5,  $p = 0.75$
4. Temperature = 1.5,  $p = 0.9$

Our diverse inference parameters were intended to create diverse automatic simplification options for our pair creators. Each parameter set was used for five of a complex sentence’s generations. We assigned the first eight prompts (i.e. the zero-shot prompts) listed in Section 3.2.1 at random during inference. We deduplicated inferences and removed inferences that left the complex sentence unchanged before proceeding with post-inference filtering and pair creation.

**Post-Inference Filtering** We applied several filters to inferences both before and during pair creation to match the SFT data distribution and simplify pair creation. The pair creators and creation process are described in the succeeding subsections.

Our initial set of post-inference filters applied to length, cosine similarity, perplexity, ROUGE scores, and intra-pair edit operations. These initial filters appear as the blue dashed vertical lines in Appendix Figures B.1 and B.2. To match the SFT data distribution, we first removed inferences more than 30 words in length and with complex-simple embedding cosine similarity less than 0.5 (embeddings were created as in Section 3.2.1). We also removed inferences with perplexity above 15 to prevent particularly unexpected simplifications from entering preference pairs. Next, we removed inferences with complex-simple ROUGE-1 F1 scores above 0.95, ROUGE-2 F1 scores above 0.9, or ROUGE-L F1 scores above 0.9. Our ROUGE filters were more relaxed than the 0.8 thresholds applied to the SFT data because we believed human pair creators could identify substantive task completion from small changes to input. The filter on ROUGE-1 F1 scores was even more relaxed because a



larger share of inferences had ROUGE-1 F1 scores above 0.9 than ROUGE-L F1 scores above 0.9 and especially ROUGE-2 F1 scores above 0.9.

The aforementioned filters were applied uniformly across all inferences. We also applied a select few filters to possible *pairings* of inferences. The first required two inferences to be separated by 3-30 word-level Levenshtein edit operations. The second required all ROUGE F1 scores between simplifications to be at or below 0.9. The chief motivation for both was a desire to avoid pairs with nearly identical simplifications, which might complicate pair annotation.

After pair creation began, we reduced simplification options even further due to the significant time investment and adverse effects on pair creator focus induced by large simplification sets. These additional ‘online’ filters, which appear as the red dotted vertical lines in Appendix Figures B.1 and B.2, were developed based on the first 1,030 pairs created by our pair creators.

Using these initial pairs we analyzed inference parameters, simplification word counts, complex-simple cosine similarities, and complex-simple ROUGE scores for both created pairs and simplification sets skipped for lack of suitable options. The results, visible in Appendix Figures B.3 and B.4, motivated the online inference filters we applied to remaining pair creation. These filters removed inferences with (1) temperature = 1.5 and p = 0.9 during generation, (2) complex-simple cosine similarity less than 0.65, and (3) complex-simple ROUGE-1 or ROUGE-L F1 score less than 0.2. The result was a 60% decrease in the global inference set and reduction in the maximum inference count displayed per complex sentence from 20 to 15. Only 24% of the first 1,030 created pairs would not have been created with these filters in place, indicating a trade-off for later pairs that was reasonable given pair creator time constraints and the abundance of remaining complex sentences. The aforementioned filters were implemented until all remaining pairs were created.

**Human Pair Creators** We recruited thirteen German-speaking human pair creators with strong backgrounds in computational linguistics research to review automatic simplifications and construct ATS pairs for preference annotation. Each pair creator was assigned a random block of 300, 600, or 1200 complex sentences and associated inferences. Creators were also provided a set of optimal pair criteria and a simple annotation tool, which we describe below.

**Criteria** We trained pair creators to search for the following four criteria when choosing automatic simplifications for a pair:

- Perfect entailment

Creators ensured the complex sentence entailed both automatic simplifications. This meant the information in the simplifications could be inferred based on the information in the complex sentence alone. Given the SFT data entailment issues discussed in Section 3.2.1, we needed this criterion to correct for the hallucinations common in our SFT model checkpoints.

We allowed an exception to perfect entailment when real-world context made any added information unambiguously true. This exception was a means of preserving in our ATS pairs the simplification strategy of adding information. The following English-language example was used to explain conditions necessary for the exception to pair creators:

- Complex: ‘Trump made a speech before his supporters angrily stormed the U.S. Capitol.’
- Simple: ‘Trump spoke. Then his supporters attacked the Capitol. They were angry *about the 2020 election.*’

While the complex sentence does not entail the clause about the 2020 election, pair creators may be aware that the only attack on the U.S. Capitol to occur after someone named Trump spoke was caused by anger over the 2020 election. Because the added information is unambiguously true, the entailment exception can be applied.

- Equal information level

Creators were asked to maintain the same amount of information from the complex sentence in both simplifications if possible. This meant that if one simplification excluded or added information, the other simplification should have ideally excluded or added the same information.

The motivation for this criterion was feedback from target group annotators in a pilot annotation session. During this pilot session, a target group annotator expressed confusion about the task when the information levels in the simplifications differed. They stated that they instead preferred to focus on the way information was written. To cater to target user needs, we established this criterion and restarted pair creation. However, we still allowed creators to create ATS pairs with differing levels of information if information equality was not possible. This is because we reasoned annotation data revealing human preferences on the threshold for information inclusion or exclusion would be valuable from a research perspective.

All pairs were created with a label indicating whether the two simplifications had an equal or unequal level of information.

- High Quality

We asked creators to select simplifications that retained the sentence’s core information while making it easier to comprehend. We also asked creators to ensure the simplifications respected the rules of the German language. The open-ended nature of text simplification made this criterion somewhat subjective; as a heuristic approach, we asked creators to select pairs that would force annotators to think for a moment before selecting their preference because both options were satisfactory.

- High Difference

Our last preference pair criterion was high difference between the two automatic simplifications. We sought high differences so preference optimization could exploit larger variations



when updating model weights. Given the similarities in entailment, information, and quality described above, we recommended creators search for differences in the domain of simplification strategy. For example, one simplification might split information while the other modifies diction. Another example might have one simplification which reorders information while the other simplifies grammar.

To clarify our priorities, we distributed the following hierarchy of criteria combinations to each pair creator:

1. Perfect entailment, equal information, high quality, high difference
2. Perfect entailment, equal information, high quality
3. Perfect entailment, high quality, high difference
4. Perfect entailment, high quality
5. Perfect entailment, medium quality

Combination 1 was our most preferred combination while Combination 5 was the least preferred combination we were willing to accept. Recognizing the threshold between ‘high’ and ‘medium’ quality was subjective, we specified that at the very least, both simplifications needed entailment (barring the aforementioned exception) and to make the complex sentence easier to comprehend in some way, however slight.

**Creation Tool** All pair creators investigated a distinct set of complex sentences with a simple Python-based pair creation tool. Upon booting up the tool, a random complex sentence is displayed. The pair creator then reads up to 20 automatic simplifications in alphabetical order and searches for two which meet our criteria. If such a pair is found, the creator enters (1) numerical indices for the selected pair and (2) a binary indicator for information equality or inequality. If the pair is not permitted due to the intra-pair ROUGE F1 ceilings described above, the tool requires the creator to try selecting a different pair. Otherwise, the pair is saved, and the tool proceeds to the next complex sentence. If no pair can be found, the pair creator may skip the current set of inferences. Up to three sets of automatic simplifications may appear per complex sentence (one for each of our winning SFT model checkpoints, unbeknownst to the pair creator). These sets appear in random order. If the creator skips all available inference sets for a complex sentence, the complex sentence is removed from consideration.

**Abandoned ATS pairs** A fourteenth pair creator created 1,800 ATS pairs with a version of the aforementioned Python script that displayed automatic translations of inferences into English.

Following creation of these 1,800 pairs, we trained an SVM model to classify inferences suitable for inclusion in an ATS pair. Unfortunately, due to concerns about translation-induced bias, the SVM was abandoned and the 1,800 pairs were excluded from HF4ATS. We did not pursue training an SVM on other creators’ pairs due to the initial SVM’s poor classification accuracy.

**Final Preference Pair Statistics** We received 3,037 preference pairs from 13 pair creators. Each pair contained a unique complex sentence from the 8,000 complex sentences we sampled from DEPLAIN - APA and LHA-APA as well as two inferences from a single model checkpoint. We did not acquire pairs for all 8,000 complex sentences due to time constraints.

Figure 3.7 shows that pair creators exhibited individual preferences for specific SFT checkpoints despite the model-blind pair creation procedure. That said, the global distribution shows less extreme concentration, with each model representing at least 28% of all pairs. This semi-equitable model representation in HF4ATS-DPO motivates our research question concerning alignment between the model being post-trained and the ATS pair source during DPO post-training.

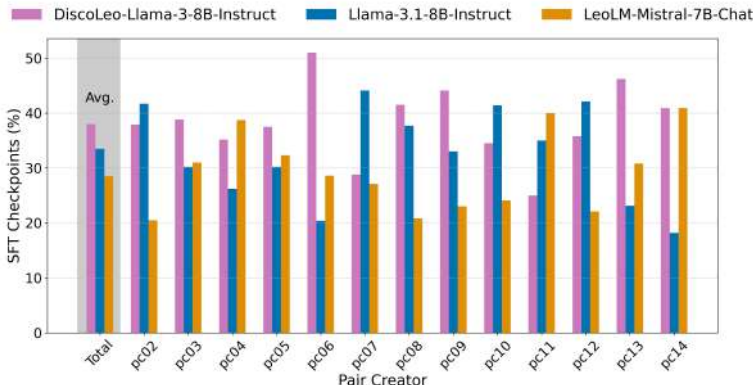


Figure 3.7: Distribution of SFT model source for created ATS pairs. This figure reflects the relative prevalence of different SFT backbone LLMs in the HF4ATS-DPO dataset and indicates that human preferences differ from the model perspective. The shaded left bar (Avg.) shows overall averages ranging from approximately 28% to 37%, with a plurality of pairs coming from the DiscoLeo-Llama model.

Figure 3.8, meanwhile, shows that pair creators indicated information equality for a large majority of our ATS pairs. This confirms that pair creators prioritized information equality during creation. At the same time, the non-negligible share of pairs without information equality provided us with the variation needed to explore information equality’s role in preference alignment efficacy.

Appendix Figures B.1 and B.2 (the same figures introduced during the description of our inference filtering) display feature distributions for inferences and inference pairs respectively. Column 1 shows distributions for all deduplicated inferences (in the former figure) and their resulting potential pairs (in the latter figure). That is, Column 1 represents the global set of possibilities for inclusion in

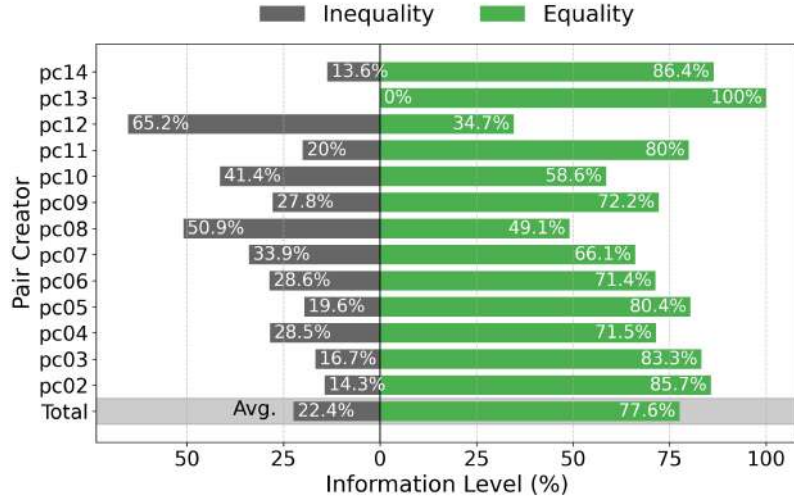


Figure 3.8: Information-level annotation in the dataset HF4ATS-DPO. This figure shows percentage of ATS preference pairs labeled by each pair creator as containing either equal or unequal information, with almost 80% of all pairs possessing information equality. This categorization reflects whether both options preserve the same content from the original input, offering insight into whether information equity plays an essential role in DPO post-training for ATS.

HF4ATS-DPO. Column 2 shows the same distributions for all remaining data after the initial offline filters were applied, and Column 3 shows the same distributions for data in the final set of 3,037 manually created pairs. Note that the online filters indicated by the vertical dotted red lines were only applied after approximately one third of pairs had been created.

The final set of manually created pairs yielded feature distributions which improved upon (by reducing skew) or were at least consistent with the global inference set’s distributions. The frequency of manually selected data also dropped significantly *before* all distributions approached the upper- or lower-bounds imposed by our filters, suggesting our filters did not seriously complicate pair creators’ search for the inference criteria we describe above. Finally, we note that our manually created pairs maintained distribution symmetry in perplexity and word count. In other words, while investigating our created ATS pairs, we did not identify any feature that might induce systemic, unwanted bias during pair annotation.

### 3.3.2 ATS Pair Annotation

All 3,037 pairs were annotated at least once by a member of two distinct groups: persons with cognitive impairments (target group annotators) and text simplification professionals (expert group annotators). Annotations took place on our custom-built web application between February and April 2025. All preference data has been anonymized.

**Web Application** We used React with JavaScript to design our web application and deployed a Flask backend API to monitor and receive annotation data from the application front-end. The web application was hosted on University of Zurich IT infrastructure in the Computational Linguistics and Linguistic Research Infrastructure groups. All annotation data was hosted on University of Zurich IT infrastructure or Switch Drive cloud services, pursuant to data retention regulations.

Our annotation tool was designed to induce as little cognitive load as possible. Upon entering our web application’s URL, an annotator enters a personalized user ID. The app then displays a set of instructions followed by an ATS pair. The austere preference pair layout with clear bounding boxes is visible in Figure 3.9; for a full view of the webpage after sign-in, see Appendix Figure D.1. For each pair, the annotator reads the two visible ATS options and identify the option which they find easier to comprehend. To indicate their preference, the annotator clicks the button labeled *Diesen Text Verstehe Ich Besser* (I understand this text better) within a simplification’s bounding box. The selected simplification then becomes highlighted in light green, as shown in Appendix Figure D.2. The annotator may proceed to more preference pairs by selecting *Weiter* (Next). He or she may submit his or her cumulative set of annotated pairs by clicking *Abschicken* (Submit), also visible in Appendix Figure D.2. This submission button can be clicked as often as desired and is visible any time the pair currently displayed has been annotated.



Figure 3.9: Partial view of annotation web application before selecting a preference.

We held pilot annotation sessions to finalize the web application’s design. These pilot sessions included one target annotator, one expert annotator, and a previous version of the web application. We generated 100 ATS pairs for annotation with random complex sentences from DEPLAIN - APA data, preliminary versions of the zero-shot prompts listed in Section 3.2.1, and ChatGPT-4o [28]. Following the pilot sessions, we made several changes to our web design. These included:

- Placing the *Abschicken* button alongside every pair rather than exclusively the final pair in an annotator’s set. This allowed annotators to submit their work regardless of their progress.
- Removing a progress bar. This prevented target group annotators from viewing other annotators’ progress and perhaps altering their annotation behavior as a result.
- Auto-scrolling the webpage to the bottom of the page (beyond the instructions and to the pref-

erence pair) after each click of *Weiter*. This streamlined annotations, especially for annotators with motor issues.

- Combining a given annotator's pairs into one continuous stream rather than compartmentalized 'buckets' accessible through an obsolete row of buttons. This streamlined annotations by removing the need to click a new button after every 25 pairs.
- Displaying the complex sentence to half of the expert annotators. This balanced the need to maintain similar annotation conditions between target and expert annotators with feedback from the pilot-session expert annotator suggesting access to the complex sentence would improve annotation quality.

Complex sentences were not shown to any target group annotators. We decided to exclude them based on (1) previous text simplification research that found persons with cognitive impairments may skim or skip long and difficult text [57] and (2) guidance from an accessibility researcher at ZHAW, who believed asking target group annotators to read complex sentences might severely complicate (and slow) the annotation task. As mentioned and motivated in our list of web design changes above, we displayed complex sentences to half of our expert annotators (denoted ea02 and ea04 in Sections 5.1 and 5.2). No annotators were informed of this difference in annotation conditions.

Based on feedback from collaborators at ZHAW, we also removed modifiable sliders intended to supplement preference annotations with length and complexity evaluations. The motivation for their removal was, again, to reduce the cognitive load imposed on annotators.

**Target Annotator Sessions** The target group annotation sessions were facilitated by our Austria-based project partner. Prior to the start of each annotator's first session, an educational coordinator (i.e., a proctor) supporting the target group participants provided web tool instructions, a consent form, an information sheet, and a questionnaire. The coordinator also demonstrated the annotation process through example tasks to familiarize the target group participants with the procedure. Each participant was provided with a tablet and unique log-in ID. Once they accessed the web tool, they annotated independently under the coordinator's supervision. Overall, we organized 15 annotation sessions, each attended by 1-10 of our 15 target group participants. The composition of annotators changed between sessions due to temporary absences, withdrawal of participation, and withdrawal of invitation (see Section 5.1 for more details on the lattermost reason). The participants had an average age of 27.4, and each was previously assessed to have a mild to moderate cognitive impairment. Target group participants were compensated at a uniform rate of 10 euros per working hour.

The University of Zurich Faculty of Philosophy Ethics Committee granted approval for our annotation sessions under Study Number 24.12.05. The information sheet, informed consent form,

instruction sheet, and questionnaire distributed during annotation sessions were written in German *Liechte Sprache* (Easy Language). These forms are included in Appendix C.

**Expert Annotator Sessions** In parallel, we recruited four native German-speaking annotators with expertise in text simplification (the expert group) to perform the same preference annotation task on the HF4ATS-DPO dataset. To reduce inter-group bias, two of the expert group participants (denoted in our results as ea01 and ea03) were unable to see each pair’s corresponding complex text, matching the target group’s annotation conditions. As stated above, to balance this inter-group bias reduction with a desire to leverage the expert group’s professional training, we displayed corresponding complex texts to the two remaining expert group participants (ea02, ea04). This latter design is illustrated in Appendix Figure D.3; the view for all other participants had the box containing *Original Text* (Original Text) removed.

Expert group participants were compensated for their annotations at an hourly rate. Ethical approval for the expert group annotations was not required. The expert annotators accessed our web application on their personal devices at times and places of their choosing. Each received an English-language video tutorial and the same German-language instruction sheet provided to our target group annotators. Among the most consistent annotators, the expert group participants completed the annotation tasks significantly faster than most target group participants; during the final evaluation annotations, expert annotators averaged 180 pairs per hour against the target group’s 60 pairs per hour.

**Sanity Check Pairs** To measure preference consistency within each annotation group and for each annotator, we introduced repeated (within-annotator consistency) and shared (within-group consistency) pairs into the annotation process. We targeted a per-participant incidence of 40-45 repeated pairs and 40-45 shared pairs, in total amounting to 10% of each expert group participant’s annotations and at least 10% of each target group participant’s annotations. Sanity check pairs were randomly sorted into each participant’s set and pairs’ constituent simplifications were presented in random order upon each appearance. These sanity check pairs were then used to calculate intra- and inter-annotator agreement (Intra-AA and Inter-AA) to measure individual preference consistency and group-level preference consistency respectively. Inter-AA was computed separately for the target group and the expert group, and only used pairs annotated by at least four annotators.

We also included complex-vs-simple pairs to determine whether annotators indeed preferred simplified text, which would indicate (1) suitable ATS pair creation and (2) successful task adherence. The number of complex-vs-simple pairs was significantly lower than intra- and inter-AA pairs for target group annotators: 10-20 pairs as opposed to 40-45. For expert annotators, we included 45 complex-vs-simple pairs. As with intra- and inter-AA data, these pairs were shuffled into annotators’ pair sets randomly. The side on which the complex sentence appeared during annotation was randomized. Complex-vs-simple pairs did not appear for the two experts who were able to view

each pair’s associated complex sentence.

We report Intra-AA, Inter-AA, and complex-vs-simple pair results in Section 5.1.

## 3.4 Direct Preference Optimization

### 3.4.1 Dataset Construction

We gathered annotations for all 3,037 ATS pairs from both target group persons and text simplification experts. Intra-AA pairs, Inter-AA pairs, complex-vs-simple pairs, and low-quality pairs all posed issues for our DPO implementation.

We first removed all complex-vs-simple sanity check pairs because they contained data that did not represent task completion. We then removed all data from target group annotators ta06, ta08, and ta09. Reasons for this data removal include annotation behavior that suggested the annotator(s) did not understand or adhere to task instructions (see Section 5.1) as well as the annotator(s) admitting to the session coordinator that they did not understand task instructions. We ensured other target group annotators re-annotated removed preference pairs to prevent ATS pair loss. Next, to remove duplicated pairs from our data, we retained the second (or latest) instance of any pair used for Intra-AA data and retained the modal preference for any pair involved in Inter-AA data. If the preferences for a pair annotated by multiple target group persons or text simplification experts exhibited an even split, we retained the preference expressed by the annotator with the highest Intra-AA. For the target group, preferences expressed by those annotators with the three lowest Intra-AA scores were not involved in the modal preference calculation when de-duplicating Inter-AA pairs.

The final filter we applied to the raw set of preference pairs relates to pair quality. While annotating pairs, one of the expert annotators temporarily documented each pair that, due to language errors, content differences, or formatting issues, complicated task completion. We identified 23 pairs from this set that failed to meet the standards we had hoped for during pair creation. Furthermore, an annotation session proctor documented seven pairs which caused similar issues for target group annotators, of which two had been separately reported by the text simplification expert. We dropped the combined 28 pairs from both target annotator data and simplification expert data. Note that the expert’s problematic pair set was compiled after approximately 500 annotations, meaning it did not consider the full set of preference pairs. At 23 out of approximately 500 pairs, this suggested a 4-5% issue rate among our preference pair dataset. The full set of pairs and associated commentary compiled by the simplification expert (which nonetheless contains some pairs that were deemed fit for inclusion) is available in Appendix E; a deeper discussion of the issues they typify is included in Section 5.1.

Using this pared down and de-duplicated set of 3,009 preference pairs, we devised a slate of preference pair subsets for DPO post-training to answer our research questions. Unless otherwise specified, each subset was used for six total post-trainings: once for target group preferences and once for expert group preferences using each of our three winning SFT checkpoints. In all cases, 80% of preference pair subset data was used for training, 10% was used for development, and 10% was reserved for test data.

The slate of preference pair subsets was as follows:

1.  $(x, y_w, y_l)_{\text{All}} = \text{All Pairs}$ . DPO on all 3,009 preference pairs.
2.  $(x, y_w, y_l)_{\text{All}_\text{I}} = \text{Information Equality Pairs}$ . DPO on all 2,335 preference pairs labeled with information equality.
3.  $(x, y_w, y_l)_{\text{LLM}_\text{I}} = \text{Model-Specific Pairs}$ . DPO on all preference pairs created with inferences from the SFT checkpoint being post-trained (1,150 for Disco Llama, 999 for Llama, and 860 for Mistral).
4.  $(x, y_w, y_l)_{\text{max. Intra-AA}} = \text{Highest Intra-Annotator Agreement Pairs}$ . DPO on all preference pairs annotated by the four target group annotators or two expert annotators with the highest intra-annotator agreement (1,276 pairs from ta04, ta05, ta10, and ta12; 1,521 pairs from ea02 and ea03).
5.  $(x, y_w, y_l)_{\text{max. Inter-AA}} = \text{Highest Inter-Annotator Agreement Pairs}$ . DPO on all preference pairs annotated by the four target group annotators or two expert annotators with the highest inter-annotator agreement (1,409 pairs from ta02, ta07, ta10, and ta11; 1,534 pairs from ea01 and ea02).
6.  $(x, y_w, y_l)_{\text{all}_{ex} \rightarrow \text{max. Intra-AA}_{tg}} = \text{Cross-Group Pairs}$ . DPO on all 3,009 expert annotator preference pairs for train data and all 1,276 preference pairs from ta04, ta05, ta10, and ta12 (the four annotators with the highest Intra-AA) for development data. We did not implement the reverse scenario (i.e. target group preferences for train data, expert group preferences for test data).

Our final evaluation is mostly concentrated on the “all” subset.

Given that these subsets differ in the number of training data points, one byproduct of their implementation is insight into the amount and type of preference data required to effectively personalize LLMs with 7-8 billion parameters. Such insights could inform more efficient planning of data annotation efforts involving target group participants.



### 3.4.2 DPO Training

We used the TRL DPOTrainer library [74] to implement DPO. Apart from reducing the training batch size to 8 from 16, we made no changes to the SFT phase training parameters described in Section 3.2.1. Our  $\beta$  parameter for Equation 2.3 was 0.1. In accordance with the DPO algorithm, prompt tokens were masked during all trainings. Once again, all trainings occurred on a single A100 GPU. Prompts were assigned at random from the set of zero-shot prompts used during SFT, as discussed in 3.2.1.

## Chapter 4

# Evaluation

### 4.1 Datasets and Model Checkpoints

Dataset	# Instances			# words		
	Train	Dev	Test	Train	Dev	Test
HF4ATS-SFT ( $\mathcal{D}_{\text{SFT}}$ )	3,600	800	800	252,285	55,208	55,852
HF4ATS-DPO ( $\mathcal{D}_{\text{DPO}}$ )	4,814	602	602	372,687	45,857	45,992

Table 4.1: **Overall statistics of the HF4ATS dataset.** We curated separate datasets for training SFT models from pre-trained LLMs and for training DPO models using preference annotations collected from either target group participants or expert group participants.

We use HF4ATS test data as well as six winning DPO checkpoints and our three winning SFT checkpoints to report final experiment results.

In Table 4.1, we showcase overall statistics for our HF4ATS dataset. As described in Section 3.2.1, 70% of the 5,200 HF4ATS-SFT ( $\mathcal{D}_{\text{SFT}}$ ) complex-simple text pairs sampled from DEPLAIN-APA were used to conduct SFT.  $\mathcal{D}_{\text{SFT}}$  development data was used to hyperparameter tune and select winning SFT checkpoints for DPO post-training, while the  $\mathcal{D}_{\text{SFT}}$  test data was reserved to evaluate winning DPO checkpoints against pre-DPO winning SFT checkpoints in Section 5.2.

HF4ATS-DPO ( $\mathcal{D}_{\text{DPO}}$ ), meanwhile, represents the 6,018 ATS preference pairs—3,009 unique pairs each annotated at least once by a target group participant and at least once by an expert group participant—used to train and evaluate our DPO implementation. 80% of  $\mathcal{D}_{\text{DPO}}$  pairs were used to post-train our DPO policy models, 10% were used to select winning DPO checkpoints based on win rates, and a final 10% were used to calculate these winning DPO checkpoints’ win rates on withheld data.

Note that the number of annotated preference pairs available in HF4ATS is larger than reported

for  $\mathcal{D}_{\text{DPO}}$  in Table 4.1. This is because pairs annotated more than once were de-duplicated, new annotation data continued to be collected after DPO post-training was implemented, and some pairs were excluded from training due to grammar or content errors. The public data release contains a field indicating whether a pair annotation was included in  $\mathcal{D}_{\text{DPO}}$ .

Pre-trained LLMs	SFT Checkpoint	DPO Checkpoint	
		Target	Expert
DiscoLeo-Llama-3-8B-Instruct	DiscoLeo-Llama-SFT-2800	DiscoLeo-Llama-DPO-2160	DiscoLeo-Llama-DPO-1080
Llama-3.1-8B-Instruct	Llama-SFT-2400	Llama-DPO-1440	Llama-DPO-1320
LeoLM-Mistral-7B-Chat	LeoLM-Mistral-SFT-1600	LeoLM-Mistral-DPO-1560	LeoLM-Mistral-DPO-2280

Table 4.2: **Model sequences from our pre-train  $\rightarrow$  SFT  $\rightarrow$  DPO pipeline** used to personalize LLM-based ATS. We reiterate that DPO checkpoints were trained separately using target and expert group annotations from HF4ATS-DPO, while SFT checkpoints were not group-specific.

Table 4.2 lists all winning model checkpoints involved in our overall training pipeline. The numbers following the SFT and DPO checkpoints indicate the number of training data points associated with each checkpoint. The DPO checkpoints refer specifically to our training variant that included all group-appropriate preference pairs in  $\mathcal{D}_{\text{DPO}}$ ; they are the “winning” DPO checkpoints because they are the checkpoints for which post-trainings on all preference pairs achieved their highest win rates on  $\mathcal{D}_{\text{DPO}}^{\text{dev}}$ . These are the only DPO checkpoints for which we implement *all* automatic evaluation and for which we implement the human evaluation described below.

## 4.2 Automatic Evaluation

Each subset of preference pairs listed in Section 3.4.1 resulted in six DPO post-trainings, one for each combination of reference SFT checkpoint (DiscoLeo-Llama-SFT-2800, Llama-SFT-2400, LeoLM-Mistral-SFT-1600) and supervision source (target or expert). To evaluate the six winning DPO checkpoints trained on all group-appropriate preference pairs, we used greedy decoding to generate inferences for all 800 complex sentences in  $\mathcal{D}_{\text{SFT}}^{\text{test}}$ , calculated the reference-based metrics SARI [76] and BERTScore [79] as well as the reference-free metric WSTF<sub>4</sub> [20], and compared these metrics to the same metrics calculated with our winning SFT checkpoints.

We then utilized  $\mathcal{D}_{\text{DPO}}^{\text{test}}$  to calculate win rates [51] for all 30 winning DPO checkpoints from our trainings. Specifically, given  $\mathcal{D}_{\text{DPO}}^{\text{test}} = \{(x_i, y_i^w, y_i^l)\}_{i=1}^N$ , where  $x$ ,  $y_w$ , and  $y_l$  denote the complex sentence, the preferred text simplification, and the dispreferred text simplification respectively, the **win rate**  $W_{y_w > y_l}$  is defined as the proportion of preference pairs for which the DPO checkpoint assigns a higher implicit reward to the preferred text simplification than the dispreferred simplification. That is,

$$W_{y_w > y_l} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{r}(x_i, y_i^w, y_i^l) > 0], \quad (4.1)$$

where  $\hat{r}(x, y_w, y_l)$  denotes the log-odds ratio computed with the policy model and reference model, as defined in Equation 2.3, and  $\mathbf{1}[\cdot]$  is the indicator function, which equals 1 if the condition holds and 0 otherwise. A win rate above 0.50 indicates that the DPO policy model more often assigns higher implicit rewards to human-preferred simplifications than dispreferred simplifications, thereby achieving closer alignment to human judgments of ATS quality.

We also report the progression of average reward margins during our DPO trainings. This corresponds to the average of the implicit reward difference seen inside the indicator function in Equation 4.1.

### 4.3 Human Evaluation

We also aimed to assess whether the target group and expert group participants indeed favored text simplifications produced by models that were personalized using their own preferences as supervision signals. For this we calculated the DPO **supremacy score**  $S_{\text{DPO} > \text{SFT}}$  for the six winning DPO checkpoints trained on all group-appropriate preference pairs. This score is defined as the proportion of text simplifications generated by the DPO checkpoint that human evaluators prefer over simplifications produced by the corresponding SFT checkpoint (i.e., the reference model). That is, if  $(y_{\text{DPO}}^{x_i}, y_{\text{SFT}}^{x_i})$  is a pairing of one DPO-checkpoint inference and one SFT-checkpoint inference, where the SFT checkpoint is the precursor of the DPO checkpoint, and  $x_i \in \mathcal{D}_{\text{DPO}}$  is the corresponding complex text used to generate the two inferences, then

$$S_{\text{DPO} > \text{SFT}} = \frac{1}{N} \sum_{i=1}^N h(x_i), \text{ where } \forall x_i \in \mathcal{D}_{\text{DPO}}^{\text{test}}, h(x_i) = \begin{cases} 1, & y_{\text{DPO}}^{x_i} > y_{\text{SFT}}^{x_i} \\ 0, & \text{otherwise.} \end{cases} \quad (4.2)$$

In this context, the successful personalization of a DPO model would be indicated by a supremacy score greater than 50%.

To compute the DPO supremacy score, for every complex sentence  $x$  in  $\mathcal{D}_{\text{DPO}}^{\text{test}}$ , we generated five text simplifications with the DPO checkpoint and five text simplifications with the corresponding SFT checkpoint using top-p sampling ( $p = 0.9$ ). We then engaged one pair creator who had previously created pairs for HF4ATS-DPO to assemble from these inferences a final set of 300 ATS pairs, 50 for each of our six winning DPO checkpoints.

The pair creator was shown a complex sentence and a procession of possible ATS pairs in randomized order without being informed which checkpoints were responsible for each inference. The creator approved or rejected pairs based on the same criteria used during initial ATS pair creation. Only those complex sentences for which the pair creator could approve one pair for all six DPO checkpoints were included in the final evaluation round with human participants.

We invited the four target group participants with the highest Intra-AA scores (i.e., ta04, ta05,

ta10, and ta12) and all four expert annotators to take part in the final human evaluation sessions. Apart from the fact that all pairs were shared within the annotation groups (albeit displayed in randomized order), annotation conditions were the same as before. Importantly, only the 150 pairs associated with the three target-group DPO checkpoints were shown to target group annotators, and only the 150 pairs associated with the three expert-group DPO checkpoints were shown to expert group annotators. Based on pairwise choices between SFT- and DPO-checkpoint-generated text simplifications, we computed DPO supremacy scores separately for each evaluator.

## Chapter 5

# Results and Discussion

### 5.1 Preference Pair Annotations

We report intra- and inter-annotator agreement (Intra-AA and Inter-AA) metrics where available in Table 5.1. There is little evidence of meaningful intra- or inter-annotator agreement among target group annotators. Intra- and inter-annotator agreement for text simplification experts, meanwhile, was typically moderate or strong.

Table 5.1: **Annotator agreement scores** measured for target and expert group participants. We find evidence of moderate or strong annotation agreement among expert group participants but no such evidence for target group participants. Note that annotators ea02 and ea04 could see the complex sentences being simplified while all other annotators could not.

(a) **Intra-annotator agreement (Intra-AA)** of target and expert group participants. We calculated Cohen’s Kappa [12] as the agreement score. NA marks those annotators for which Intra-AA data is not available; ta06, ta08, and ta09 data was not included in DPO post-training while ta13 and ta15 could not attend sessions with Intra-AA pairs.

Target						Expert	
id	$\kappa$	id	$\kappa$	id	$\kappa$	id	$\kappa$
ta01	-0.037	ta06	NA	ta11	0.063	ea01	0.420
ta02	0.040	ta07	-0.045	ta12	0.155	ea02	0.755
ta03	-0.026	ta08	NA	ta13	NA	ea03	0.745
ta04	0.168	ta09	NA	ta14	0.008	ea04	0.376
ta05	0.300	ta10	0.065	ta15	NA		

(b) **Inter-annotator agreement (Inter-AA)** of target and expert group participants. We calculated Krippendorff’s Alpha [35] as the agreement score. We report Inter-AA scores for pairs annotated by at least four annotators, stratified by the generating SFT checkpoint.

SFT Checkpoint	$\alpha$	
	Target	Expert
DiscoLeo-Llama-SFT-2800	0.019	0.324
Llama-SFT-2400	0.003	0.248
LeoLM-Mistral-SFT-1600	-0.016	0.536

One potential reason for the discrepancy in Intra-AA between target and expert groups is that expert annotators were more likely to encounter repeated pairs within a short time frame. All expert annotators annotated their data within the span of a week, often at such a pace that they encountered

repeated pairs on the same day. Target annotators, meanwhile, annotated on a multi-week time frame, resulting in longer gaps between sightings of repeated pairs.

Regarding Inter-AA, it may be that expert annotators exhibited greater agreement than target annotators because expert annotators relied on shared professional knowledge of the text simplification task that was likely not shared by target annotators. Furthermore, Inter-AA data for the target group factored in up to eleven annotators compared to just four for the expert group, with probable diversity in the type and severity of target group cognitive impairments likely yielding diversity in innate preferences as well.

One final reason for target users' low or negative Intra- and Inter-AA scores may have been that they did not understand or adhere to task instructions. This is evident upon review of their annotation behavior: as shown in Figure 5.1, several target annotators appear to have chosen a preferred side (left or right) and almost always indicated preference for that side. Target annotator ta14, for example, heavily favored the right-hand side ATS. We tried reconciling this situation in different ways: by reminding such annotators about the importance of reading both simplifications before selecting an option (ta10); by neglecting to invite such annotators to succeeding sessions (ta06); by allowing such annotators to remain until all pairs were annotated, then using leftover resources to re-annotate as many of their pairs as possible (ta14). Importantly, this metric does not discount or validate an annotator's annotations outright. Consider that ta05, who selected the left-hand ATS only 20% of the time, had the highest Intra-AA among all target annotators. Nonetheless, the more consistent expert group exhibited relative indifference to ATS positioning, suggesting that some target group annotators' positional preferences were indicative of low task adherence.

We note that the presentation and quality of some ATS pairs may have complicated task adherence even for those annotators who understood our instructions. For example, our web application displayed some words with hyphenation, splitting longer words across separate lines. This unintentional effect was an artifact of our web app design rather than ATS content. The presence of hyphenation may have influenced annotator behavior along a dimension we did not wish to study, and because we did not explicitly ask annotators to ignore hyphenation, its presence may have biased annotators in different ways, causing inter-annotator agreement to suffer. The presence of language mistakes (e.g. grammar errors) in some ATS pairs despite our pair creation criteria may have had a similar effect; some annotators might have punished language mistakes heavily while others did not. These issues are present in the document of problematic pairs compiled by an expert group annotator, which we provide in Appendix E.

Separate from intra- and inter-annotator agreement, we include complex-vs-simple sanity check annotation results in Appendix Table E1, which suggest no universal preference for or against simplified text (as opposed to complex text) in either the expert or target group. Because most target annotators preferred simplified text with a rate at or close to 50% based on 10-15 complex-vs-simple pairs, none were found to prefer or disprefer simplified text with evidence of statistical significance. Meanwhile, after 45 complex-vs-simple sanity check annotations, one of the expert annotators

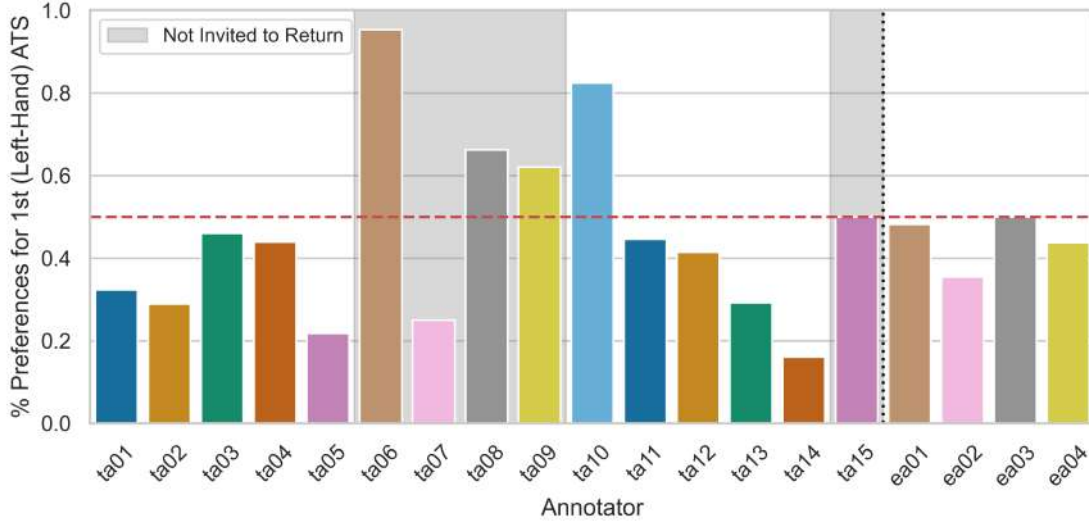


Figure 5.1: Preference rate for the left-hand option by user. Annotators such as ta06 exhibited an overwhelming preference for one side, suggesting they did not understand or adhere to task instructions. Apart from concentrating preferences on one side, annotators may not have been invited to return for other reasons (e.g. admitting they struggled to understand the task).

indicated preference for simplified text at a rate of 80% (statistically significant with  $p < 0.01$ ) while the other did not prefer or disprefer simplified text.

One reason annotators do not widely express preferences for simplified text may stem from the characteristics of the DEPLAIN - APA. The complex-simple data instances in DEPLAIN - APA combine sentences from B1-classified *articles* with sentences from A2-classified *articles*. That is, the language classifications apply to each sentence’s source article rather than the sentences themselves. This could result in “complex-simple” pairs composed of sentences that do not exhibit the expected transition in reading comprehension difficulty (the examples shown under Data Filtering in Section 3.2.1 serve as suitable examples despite being introduced to exemplify intra-pair similarity or lack of entailment). Using such complex-simple pairs for SFT might have attenuated the language models’ proclivity for simplifying text. Nonetheless, we have statistically significant evidence that the expert annotator with higher intra-annotation agreement prefers simplified text, validating our research approach.

## 5.2 DPO

We now summarize our main findings with respect to the RQs we proposed in Section 2.



### 5.2.1 Quality Assessment of Generated ATS

**RQ1** Can DPO post-training with pairwise human preferences further improve the quality of ATS, as measured by automatic evaluation metrics?

In Table 5.2, we present an automatic quality assessment of ATS outputs generated by our SFT and DPO checkpoints, evaluated using both reference-based (SARI, BERTScore) and reference-free (WSTF<sub>4</sub>, Win Rate) metrics. The best performance on SARI and WSTF<sub>4</sub> was achieved with the DPO checkpoints, whereas the highest BERTScore was obtained using the SFT checkpoints.

The decreased WSTF<sub>4</sub> scores reported by the DPO checkpoints show that DPO post-training benefits ATS readability in general, regardless of the backbone LLM or preference source. Meanwhile, DPO checkpoints have consistently lower BERTScores than the corresponding SFT checkpoints across all LLMs and both preference sources, suggesting that DPO post-training has caused a certain level of semantic drift. While the effect on simplification faithfulness (as measured by SARI) is not uniform, DPO checkpoints trained with expert group preferences—such as Llama-DPO-1320 or LeoLM-Mistral-DPO-2280—tend to maintain or recover baseline SARI scores. This contrasts with DPO training on target group preferences, where decreases in simplification faithfulness were observed in two target DPO checkpoints. Importantly, expert-supervised models uniformly achieve higher win rates than their corresponding target-supervised models, emphasizing greater preference consistency among expert group annotations. This may explain the two supervision sources’ differing impacts on readability and faithfulness. The findings above point to a trade-off between simplification strength and meaning preservation, with the consistency of supervision playing a key role in how well DPO models balance the two.

Recent studies have revealed several core limitations of standard DPO post-training. These include a tendency to overfit to sparse or noisy preference signals [18], catastrophic forgetting in continual learning settings [49], and the potential to undermine generalization and robustness in LLMs [27]. In our study, these limitations may help explain the observed decline in DPO models’ faithfulness with respect to the DEPLAIN data in  $\mathcal{D}_{\text{SFT}}^{\text{test}}$ . Nonetheless, our results highlight the critical role of preference consistency in the effectiveness of DPO for personalized ATS modeling. The win rates in Table 5.2 as well as Inter- and Intra-AA scores in Table 5.1 indicate that target group preferences are more diverse or inconsistent than expert group preferences. It might be the case that offline LLM alignment methods such as DPO, which lack explicit reward modeling, are suboptimal for capturing nuanced preferences over text simplifications when trained with such data.

**Findings** Overall, DPO post-training improves ATS in terms of text readability (as evidenced by WSTF<sub>4</sub> scores) and, in some cases, N-gram preservation (as evidenced by SARI scores). However, despite evidence that expert group supervision signals can result in closer alignment of their preferences, models trained on either expert or target group supervision exhibit loss in meaning preservation (as evidenced by BERTScore). The scale of these effects depends on the source of the supervision

Checkpoint	Reference-based Metrics		Reference-free Metrics	
SFT Baselines	SARI	BERTScore	WSTF <sub>4</sub>	Win Rate
DiscoLeo-Llama-SFT-2800	<b>46.22</b> $\pm$ 13.47	<b>0.9049</b> $\pm$ 0.054	6.515 $\pm$ 3.24	-
Llama-SFT-2400	45.94 $\pm$ 13.52	<b>0.8865</b> $\pm$ 0.054	5.852 $\pm$ 2.90	-
LeoLM-Mistral-SFT-1600	44.55 $\pm$ 13.95	<b>0.9054</b> $\pm$ 0.056	6.207 $\pm$ 3.55	-
DPO Target	SARI	BERTScore	WSTF <sub>4</sub>	Win Rate
DiscoLeo-Llama-DPO-2160	44.41 $\pm$ 11.60	0.7854 $\pm$ 0.081	6.194 $\pm$ 2.17	0.5211
Llama-DPO-1440	46.11 $\pm$ 11.60	0.8756 $\pm$ 0.055	5.796 $\pm$ 2.56	0.5145
LeoLM-Mistral-DPO-1560	43.73 $\pm$ 13.36	0.7781 $\pm$ 0.113	5.683 $\pm$ 2.80	0.4382
DPO Expert	SARI	BERTScore	WSTF <sub>4</sub>	Win Rate
DiscoLeo-Llama-DPO-1080	42.50 $\pm$ 13.28	0.7814 $\pm$ 0.059	4.031 $\pm$ 2.68	0.6118
Llama-DPO-1320	46.45 $\pm$ 12.25	0.8441 $\pm$ 0.052	4.676 $\pm$ 2.59	0.6099
LeoLM-Mistral-DPO-2280	44.92 $\pm$ 14.03	0.8340 $\pm$ 0.082	4.802 $\pm$ 2.94	0.6118

Table 5.2: **Automatic assessment of ATS quality** for inferences from both SFT and DPO checkpoints on the SFT test set  $\mathcal{D}_{\text{SFT}}^{\text{test}}$  (columns 1-3) and DPO test set  $\mathcal{D}_{\text{DPO}}^{\text{test}}$  (column 4). We report the mean scores and standard deviations (in brackets) for each metric. The same color scheme is used for win rates (values above vs. below 0.50). DPO checkpoints in the table were trained with all annotated data for the given supervision source (expert or target). The relatively high standard deviations for SARI reflect the fact that only one reference per test instance is available. Win rates are computed based on annotator preferences within the respective supervisory groups (target or expert).

signals—target group participants vs. expert group participants—as well as the backbone model, i.e. DiscoLeo-Llama-3-8B-Instruct, Llama-3.1-8B-Instruct, or LeoLM-Mistral-7B-Chat.

### 5.2.2 Impact of Individual Factors on DPO Post-training

**RQ2** To what extent to preference inconsistency, information equality, LLM backbone, and preference source influence the effectiveness of DPO post-training?

Table 5.3 presents the win rates of DPO checkpoints trained on different subsets of  $\mathcal{D}_{\text{DPO}}^{\text{train}}$ , retained because they exhibited the maximum win rate on their subset of  $\mathcal{D}_{\text{DPO}}^{\text{dev}}$  during training, and evaluated on corresponding subsets of  $\mathcal{D}_{\text{DPO}}^{\text{test}}$ . We observe that DPO models trained on subsets reflecting higher consistency of human preferences, such as those with maximized Intra- or Inter-AA scores, leads to an almost uniform improvement in win rates across all models and both preference sources. For example, the target group LeoLM-Mistral-DPO model achieves a substantial 31.93% boost to win rate for training on the Intra-AA subset relative to the baseline DPO training on all preference pairs. Meanwhile, subsets focused on human perception or model involvement variables, such as pairs labeled with information equity (all<sub>=</sub>) or pairs with the same LLM backbone as the model being post-trained (LLM<sub>=</sub>), exhibit a less consistent and frequently negative impact on performance. For instance, the win rate of the target group DiscoLeo-Llama-DPO model drops by

DPO Checkpoint	all	all=	LLM=	max. Intra-AA	max. Inter-AA
<b>Target</b>	<b>Baseline</b>	<b>Subsets of HF4ATS-DPO training data</b>			
DiscoLeo-Llama-DPO	0.5211	0.4708 (9.65% ↓)	0.4861 (6.72% ↓)	0.5078 (2.55% ↓)	<b>0.5431</b> (4.22% ↑)
Llama-DPO	0.5145	0.4833 (6.06% ↓)	0.5385 (4.66% ↑)	<b>0.6094</b> (18.45% ↑)	0.5153 (0.16% ↑)
LeoLM-Mistral-DPO	0.4382	0.4625 (5.55% ↑)	0.4848 (10.63% ↑)	<b>0.5781</b> (31.93% ↑)	0.4917 (12.21% ↑)
<b>Expert</b>	<b>Baseline</b>	<b>Subsets of HF4ATS-DPO training data</b>			
DiscoLeo-Llama-DPO	0.6118	0.6333 (3.51% ↑)	0.5111 (16.46% ↓)	0.6118 (0.00% =)	<b>0.6438</b> (5.23% ↑)
Llama-DPO	0.6099	0.5833 (4.36% ↓)	<b>0.6538</b> (7.20% ↑)	0.6382 (4.64% ↑)	0.6313 (3.51% ↑)
LeoLM-Mistral-DPO	0.6118	0.6125 (0.11% ↑)	0.5871 (4.04% ↓)	<b>0.6776</b> (10.76% ↑)	0.5625 (8.06% ↓)

Table 5.3: **Win rates** ( $W_{y_w > y_l}$ ) **on corresponding  $\mathcal{D}_{\text{DPO}}^{\text{test}}$  subsets** for the winning DPO checkpoints trained on different subsets of  $\mathcal{D}_{\text{DPO}}^{\text{train}}$ , as described in Section 3.4. Note that the number of training instances is not uniform across the reported DPO checkpoints. To avoid potential confusion, we omit the number of training instances per subset. For each subset, we saved a DPO checkpoint every 120 training instances and retained the checkpoint with the highest win rate on the corresponding  $\mathcal{D}_{\text{DPO}}^{\text{dev}}$  subset. The performance changes in brackets indicate differences relative to the results in the second column *all* (i.e., the checkpoints indicated by the circles (o) in Figure 5.2).

9.65% under the all= subset. Every DPO model trained with expert group supervision has a higher win rate than the target-group-supervised model with the same backbone and training pair subset. Additionally, no expert-group DPO model’s win rate dips below 0.50—a threshold indicating a DPO-induced drift *away* from human preferences—and win rate shifts across the different training pair subsets tend to be lower for the expert-group models than the target-group models. These results suggest that preference consistency is a stronger driver of DPO effectiveness than model-specific or perception-related signals, particularly when the goal is to robustly model human preferences. We emphasize that the benefits of preference consistency are not limited to expert group participants; consider that the aforementioned win rate boost for LeoLM-Mistral-DPO under maximum Intra-AA target group supervision represents the largest boost across all of our experiments.

To examine how different LLMs behave during DPO post-training, at every 120 training instances we visualize win rates and implicit reward margins for  $\mathcal{D}_{\text{DPO}}^{\text{dev}}$  pairs in Figures 5.2 and 5.3 respectively. In the lower-right panels of both figures, we also include results from training on all *expert* group annotations and evaluating on *target* group annotations from the maximum Intra-AA  $\mathcal{D}_{\text{DPO}}^{\text{dev}}$  subset. This setup allows us to assess whether models trained with expert preferences—which are often much easier to collect in practice—can effectively cater to the preferences of target group persons.

We observe that across most settings, win rates on expert preferences (dashed lines) exhibit more stable and consistent improvements as the number of training instances increases. Win rates on target preferences (solid lines), meanwhile, stabilize at lower levels and only slightly above the 50% border line, likely reflecting the noisier or less consistent decision-making process among target group participants. The relatively small implicit reward margins for target group trainings suggest target group hesitation in distinguishing between preferred and dispreferred ATS outputs may have contributed to this inconsistency. Notably, in the lower-right panel, where models are trained on expert group preferences and evaluated on the most consistent subset of target group

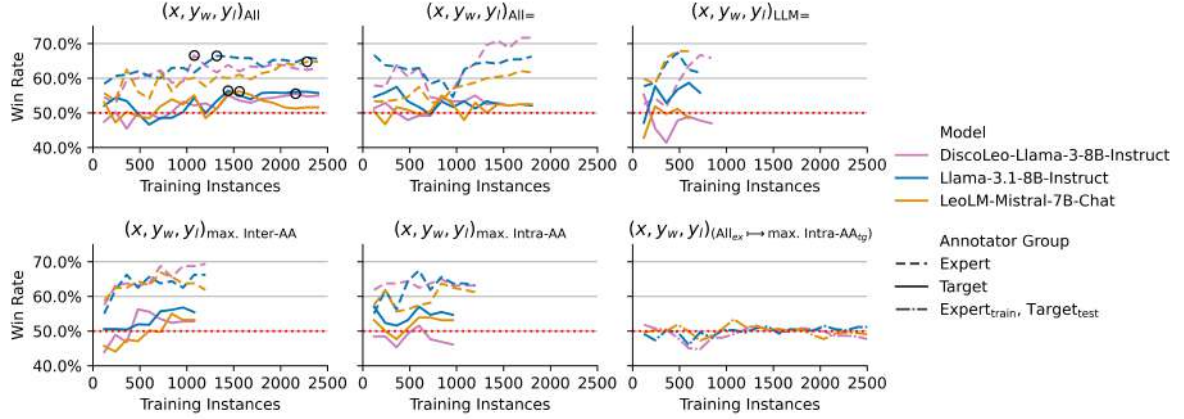


Figure 5.2: **Win rates on development sets during training** across different subsets of the HF4ATS-DPO training data. Development sets amount to 10% of maximum training instances in all cases. Circles (o) in the upper left figure indicate the DPO checkpoint with the highest win rate on  $\mathcal{D}_{\text{DPO}}^{\text{dev}}$ . The bottom-right panel shows results from training on all expert group data and testing on target group data for the four target annotators with the highest Intra-AA scores.

preferences, the win rates oscillate around the 50% threshold, suggesting limited generalization from expert preferences to target preferences. These findings underscore the importance of both preference quality and consistency in DPO post-training outcomes and suggest that DPO may not be the most effective LLM alignment method when the goal is to personalize ATS systems for target groups persons with diverse subjective preferences.

**Findings** Factors that directly reflect the consistency of human preference signals—such as intra- or inter-annotator agreement—tend to have a more positive and reliable impact on win rates. By contrast, human-perceived information equity levels in training pairs or consistency between the model being post-trained and the pair-generating model show less prominent influence, suggesting they are less reliable factors of improved preference alignment.

### 5.2.3 Personalization Success Rates of Target and Expert Models

**RQ3** Can DPO post-training enable the successful group-level personalization of ATS models despite these uncertainties?

Table 5.4 presents DPO supremacy scores for our six winning DPO checkpoints at the individual annotator level. We find that three of four target annotators have a clear preference for SFT model inferences over DPO model inferences. Furthermore, there is limited consistency among target group evaluators: ta04 and ta05 are most forgiving to LeoLM-Mistral-DPO-1560, ta12 is most forgiving to Llama-DPO-1440, and ta10 is the only annotator to indicate preference for DPO inferences.

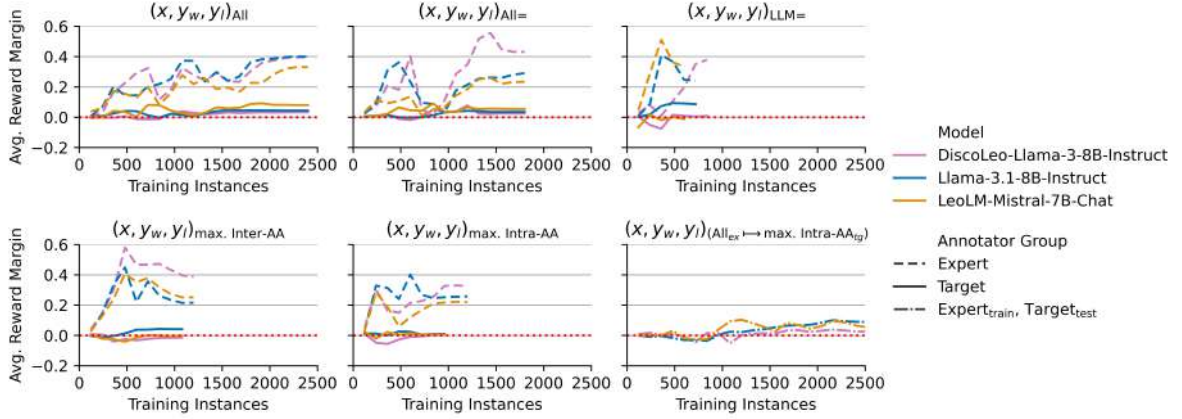


Figure 5.3: **Average reward margins with respect to the number of training instances** from different subsets of the HF4ATS-DPO training data, evaluated on the corresponding development sets.

In contrast to the target evaluators, three of four expert annotators exhibit a preference for DPO checkpoint inferences across all model backbones, and DiscoLeo-Llama-DPO-1080 performs best relative to SFT for all four annotators, typically by a sizable margin.

To verify group-level personalization, we conducted one-sided binomial tests<sup>1</sup> for each model at the evaluation group level. Assuming each pair was evaluated independently, we defined the group-level preference for each test pair as the majority vote among the evaluators (tied pairs were assigned randomly). Our goal was to determine whether, across all 50 DPO supremacy test pairs, there was a statistically significant collective preference for ATS outputs generated by the DPO checkpoints. The asterisks in Table 5.4 indicate which models had a group-level DPO supremacy greater than 0.50 with a  $p$ -value less than 0.05. For the expert group, we indicate results for tests both including and excluding the outlier ea02.

Among expert evaluators, the two German-tuned backbones produced DPO checkpoints with DPO supremacy above 0.50 at the group level. No target group DPO checkpoint demonstrates DPO supremacy with statistical significance. The greater preference consistency among expert evaluators that manifested in Table 5.1 as well as Figures 5.2 and 5.3 evidently led to a DPO post-training that was able to anchor collective expert group preferences to a certain degree. Whether due to greater diversity in innate preferences, lack of simplification training, or barriers to annotation task adherence, collective target group preferences were not reflected by DPO post-trainings in the same way.

It should also be noted that unlike for the expert group checkpoints, the target group DPO checkpoints in Table 5.4 were trained on data which included annotations from eight additional target users with lower intra-annotation agreement than ta04, ta05, ta10, and ta12. It is therefore

<sup>1</sup>We used the Python package SciPy, BSD-3-Clause license, available at <https://github.com/scipy/scipy>, to run the binomial tests.

SFT Checkpoint	DPO Checkpoint	DPO Supremacy Score			
Baseline	Target	ta04	ta05	ta10	ta12
DiscoLeo-Llama-SFT-2800	DiscoLeo-Llama-DPO-2160	0.36	0.40	0.56	0.46
Llama-SFT-2400	Llama-DPO-1440	0.40	0.30	0.38	0.50
LeoLM-Mistral-SFT-1600	LeoLM-Mistral-DPO-1560	0.42	0.48	0.58	0.40
SFT Checkpoint	DPO Checkpoint	DPO Supremacy Score			
Baseline	Expert	ea01	ea02	ea03	ea04
DiscoLeo-Llama-SFT-2800	DiscoLeo-Llama-DPO-1080 <sup>*,**</sup>	0.74	0.46	0.68	0.72
Llama-SFT-2400	Llama-DPO-1320	0.60	0.30	0.54	0.56
LeoLM-Mistral-SFT-1600	LeoLM-Mistral-DPO-2280 <sup>*</sup>	0.68	0.44	0.52	0.56

Table 5.4: **DPO supremacy scores measured with the best SFT and DPO checkpoints by LLM backbone.** We report the average proportion of text simplifications produced by DPO checkpoints that are preferred by human evaluators over simplifications produced by corresponding SFT checkpoints. For the target evaluator group, we report human evaluation results based on the four evaluators with the highest Intra-AA scores during the preference annotation stage. Asterisks indicate whether a group-level binomial test on evaluation preference with majority voting indicates DPO supremacy for a given model at the 0.05 significance level. One asterisk indicates significance when ea02 preferences are excluded and two asterisks indicate significance when ea02 preferences are included.

possible that human evaluation of a DPO checkpoint trained on a different subset of our target group preference pairs would better capture group-level preferences.

Given the impact seen in Table 5.4 of target group preference inconsistency, we now reflect on the effectiveness of DPO for ATS preference alignment for persons with cognitive impairments. During the HF4ATS-DPO annotation stage, our target group annotators occasionally encountered pairs they had trouble annotating due to the similarity or shared suitability of the two ATS options. DPO, with its contrastive, distributional learning objective, is capable of exploiting supervision from these low-signal pairs provided annotations are gathered at scale. However, gathering data at scale is a high-resource endeavor that is uniquely encumbered in accessibility settings by the need to manage the cognitive load imposed on target group persons. This means the recurring request for them to read and compare competing simplifications may work against the need for numerous annotations.

As an alternative framework, we propose exploring preference alignment methods that reduce the cognitive load on annotators, namely KTO [16]. Target annotators would only need to label a standalone simplification as desirable or undesirable to implement KTO, thus removing comparison entirely. Separately, ATS preference alignment would likely benefit from KTO’s ability to better incorporate loss aversion and other human cognitive biases due to its inference utility maximization objective. This suggests some dividends could be achieved by simply splitting HF4ATS-DPO preference pairs and using the restructured data to implement KTO, an approach we leave for future research.

**Findings** Personalization under DPO can be successful at both the individual level and group level using group-wide preferences. That said, outcomes are sensitive to barriers to preference consistency.



## Chapter 6

# Conclusion and Future Work

### 6.1 Conclusion

In this work, we studied the effectiveness of direct preference optimization (DPO) for personalizing LLM-based automatic text simplification (ATS) models to better reflect the preferences of persons with cognitive impairments. We developed a lightweight and accessible web application for collecting pairwise human preferences from both target users and text simplification experts to achieve this goal. We introduced HF4ATS, the first and largest German-language ATS dataset combining preference annotations from both target and expert group. Using a standard pre-train  $\rightarrow$  SFT  $\rightarrow$  DPO pipeline, we trained and analyzed models on various subsets of this dataset, systematically investigating how preference consistency, preference source, and LLM engagement impact personalization outcomes. Our findings indicate that DPO can (1) achieve ATS personalization, (2) gear ATS toward human preferences, and (3) improve ATS readability at the cost of semantic drift, but that consistency in supervisory sources is paramount to these effects.

### 6.2 Limitations and Future Work

Language level characteristics of DEPLAIN - APA data led to several limitations in this work. All text in DEPLAIN - APA is classified as A2 or B1, meaning its “complex” sentences may already be comprehensible to target users. Additionally, because it was compiled at the article level, DEPLAIN - APA complex sentences do not always entail simple sentences. These factors led our SFT checkpoints to hallucinate context and reward small changes to input during inference. Future researchers could improve upon our pipeline by compiling a new German manually-aligned complex-simple sentence simplification dataset for SFT that (1) simplifies complex sentences from above the B1 level and (2) enforces entailment among simplifications. Alternatively, future researchers could explore a



preference optimization pipeline that excludes SFT altogether.

The metadata and quality of our ATS pairs also introduced DPO post-training limitations. Regarding metadata, we asked pair creators to specify whether an ATS pair contained equal or unequal information, but we did not ask whether the pair exhibited high difference, high quality, or the entailment exception. This lack of granularity prevented us from exploring a wider array of preference pair subsets, including one that excluded lower quality pairs from training. Indeed, the knowledge that we were willing to accept pairs of “medium quality” provided they respected entailment may have led to the language errors seen in an estimated 4-5% of pairs. Asking pair creators for more information during creation could have helped identify such pairs, removing the bias introduced by correctness-based (rather than comprehension-based) annotation behavior. Future research could address pair quality concerns by using larger language models and directly generating ATS *pairs* rather than inferences for human review. At higher model capacities, language errors would be less likely, and pair creation criteria could be encoded directly into the prompts.

Separately, hyphenation in our displayed ATS pairs highlighted a fundamental challenge to aligning LLMs with human preferences: methods like DPO rely on consistency in supervision signals that is often difficult to achieve, particularly when preference data is collected from persons with cognitive impairments. There is a pressing need for improved HCI approaches to eliciting preferences, including adaptive and accessible tools that support our target users in particular to provide consistent, accurate, and potentially detailed feedback.

One final limitation concerns the usage of DPO itself. As mentioned in Section 5.2, collecting preferences at the scale required for successful DPO post-training is difficult when another chief priority is limiting the cognitive load imposed on annotators. This limitation calls for lightweight personalization strategies to leverage small, high-quality, consistent human preference data. Alternatively, research can explore new alignment objectives that are robust to uncertainty and variability in human preferences. More broadly, we advocate for inclusive AI development that continues to center the input of persons with impairments, not merely as end-users or evaluators, but as active co-creators throughout the research and implementation process.

# Bibliography

- [1] Sweta Agrawal and Marine Carpuat. “Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 432–448.
- [2] Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. “Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. 2023, pp. 1147–1158.
- [3] Dennis Aumiller and Michael Gertz. “Klexikon: A German Dataset for Joint Summarization and Simplification”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022, pp. 2693–2701.
- [4] Nguyen Bach, Qin Gao, Stephan Vogel, and Alex Waibel. “TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training”. In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. 2011, pp. 474–482.
- [5] Richard Bamberger and Erich Vanacek. *Lesen-Verstehen-Lernen-Schreiben*. Diesterweg, 1984.
- [6] Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. “A Corpus for Automatic Readability Assessment and Text Simplification of German”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 3302–3311.
- [7] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. “Power to the People? Opportunities and Challenges for Participatory AI”. In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2022, pp. 1–8.
- [8] Ralph Allan Bradley and Milton E Terry. “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons”. In: *Biometrika* 39.3/4 (1952), pp. 324–345.
- [9] Luisa Carrer, Andreas Säuberli, Martin Kappus, and Sarah Ebling. “Towards Holistic Human Evaluation of Automatic Text Simplification”. In: *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*. 2024, pp. 71–80.
- [10] Andrew Cashin, Julia Morphet, Nathan J Wilson, and Amy Pracilio. “Barriers to Communication with People with Developmental Disabilities: A Reflexive Thematic Analysis”. In: *Nursing & health sciences* 26.1 (2024), e13103.

- [11] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. “Deep Reinforcement Learning from Human Preferences”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 4302–4310.
- [12] Jacob Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.
- [13] Deutsches Institut für Normung e.V. (DIN). *E DIN SPEC 33429:2023-04 – Empfehlungen für Deutsche Leichte Sprache*. Accessed: 2025-02-10. 2023. URL: <https://www.din.de/de/mitwirken/normenausschuesse/naerg/e-din-spec-33429-2023-04-empfehlungen-fuer-deutsche-leichte-sprache--901210>.
- [14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. “The LLaMA 3 Herd of Models”. In: *arXiv preprint arXiv:2407.21783* (2024).
- [15] Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. “Automatic Text Simplification for German”. In: *Frontiers in Communication* 7 (2022), p. 706718.
- [16] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. “Model Alignment as Prospect Theoretic Optimization”. In: *Forty-first International Conference on Machine Learning*. 2024.
- [17] Federal Ministry of Social Affairs, Health, Care and Consumer Protection (BMSGPK). “National Action Plan on Disability 2022–2030”. In: *Vienna: BMSGPK* (2022).
- [18] Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. “Robust Preference Optimization Through Reward Model Distillation”. In: *arXiv preprint arXiv:2405.19316* (2024).
- [19] Lukas Fischer, Yingqiang Gao, Alexa Lintner, and Sarah Ebling. “SwissADT: An Audio Description Translation System for Swiss Languages”. In: *arXiv preprint arXiv:2411.14967* (2024).
- [20] Rudolph Flesch. “A New Readability Yardstick”. In: *Journal of applied psychology* 32.3 (1948), p. 221.
- [21] Yingqiang Gao, Lukas Fischer, Alexa Lintner, and Sarah Ebling. “Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies”. In: *arXiv preprint arXiv:2410.08860* (2024).
- [22] Annette Rios Gonzales, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. “A New Dataset and Efficient Baselines for Document-level Text Simplification in German”. In: *Proceedings of the Third Workshop on New Frontiers in Summarization*. 2021, pp. 152–161.
- [23] Silvia Hansen-Schirra, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, Silvana Deilen, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. “Intralingual Translation into Easy Language—or How to Reduce Cognitive Processing Costs”. In: *Easy Language Research: Text and User Perspectives*. Berlin: Frank & Timme (2020), pp. 197–225.

- [24] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. “Towards a Unified View of Parameter-Efficient Transfer Learning”. In: *Proceedings of the Tenth International Conference on Learning Representations*. 2022.
- [25] Freya Hewett, Hadi Asghari, and Manfred Stede. “Elaborative Simplification for German-language Texts”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2024, pp. 29–39.
- [26] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *Proceedings of the Tenth International Conference on Learning Representations*. 2022.
- [27] Xiangkun Hu, Tong He, and David Wipf. “New Desiderata for Direct Preference Optimization”. In: *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.
- [28] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. “GPT-4o system card”. In: *arXiv preprint arXiv:2410.21276* (2024).
- [29] Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. “Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback”. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. 2024.
- [30] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).
- [31] Tannon Kew and Sarah Ebling. “Target-level Sentence Simplification as Controlled Paraphrasing”. In: *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. 2022, pp. 28–42.
- [32] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. “Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel”. In: (1975).
- [33] David Klaper, Sarah Ebling, and Martin Volk. “Building a German/Simple German Parallel Corpus for Automatic Text Simplification”. In: *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. 2013, pp. 11–19.
- [34] Lars Klöser, Mika Beele, Jan-Niklas Schagen, and Bodo Kraft. “German Text Simplification: Finetuning Large Language Models with Semi-Synthetic Data”. In: *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*. 2024, pp. 63–72.
- [35] Klaus Krippendorff. “Reliability in Content Analysis: Some Common Misconceptions and Recommendations”. In: *Human communication research* 30.3 (2004), pp. 411–433.

- [36] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7871–7880.
- [37] Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. “A Survey of Personalized Large Language Models: Progress and Future Directions”. In: *arXiv preprint arXiv:2502.11528* (2025).
- [38] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. “Multilingual Denoising Pre-training for Neural Machine Translation”. In: (2020), pp. 726–742.
- [39] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*. 2019.
- [40] Ilya Loshchilov and Frank Hutter. “SGDR: Stochastic Gradient Descent with Warm Restarts”. In: *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022)*. 2022.
- [41] Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. “LENS: A Learnable Evaluation Metric for Text Simplification”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 16383–16408.
- [42] Philip May. *Cross English & German RoBERTa for Sentence Embeddings*. <https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer>. Accessed: 2025-02-06. 2025.
- [43] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. “Human-in-the-loop Machine Learning: A State of the Art”. In: *Artificial Intelligence Review* 56.4 (2023), pp. 3005–3054.
- [44] Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. “Considerations for Meaningful Sign Language Machine Translation Based on Glosses”. In: *The 61st Annual Meeting Of The Association For Computational Linguistics*. 2023.
- [45] Akifumi Nakamachi, Tomoyuki Kajiwar, and Yuki Arase. “Text Simplification with Reinforcement Learning using Supervised Rewards on Grammaticality, Meaning Preservation, and Simplicity”. In: *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing: Student research workshop*. 2020, pp. 153–159.
- [46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Vol. 35. 2022, pp. 27730–27744.

- [47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.
- [48] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. “The Ultimate Guide to Fine-tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities”. In: *arXiv preprint arXiv:2408.13296* (2024).
- [49] Biqing Qi, Pengfei Li, Fangyuan Li, Junqi Gao, Kaiyan Zhang, and Bowen Zhou. “Online DPO: Online Direct Preference Optimization with Fast-slow Chasing”. In: *arXiv preprint arXiv:2406.05534* (2024).
- [50] Jipeng Qiang and Xindong Wu. “Unsupervised Statistical Text Simplification”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.4 (2019), pp. 1802–1806.
- [51] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. “Direct Preference Optimization: Your Language Model is Secretly A Reward Model”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023, pp. 53728–53741.
- [52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. “Exploring the Limits of Transfer Learning with A Unified Text-to-text Transformer”. In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.
- [53] Anja Ryser, Yingqiang Gao, and Sarah Ebling. “Digitally Supported Analysis of Spontaneous Speech (DigiSpon): Benchmarking NLP-Supported Language Sample Analysis of Swiss Children’s Speech”. In: *arXiv preprint arXiv:2504.00780* (2025).
- [54] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. “Making it Simplext: Implementation and Evaluation of A Text Simplification System for Spanish”. In: *ACM Transactions on Accessible Computing (TACCESS)* 6.4 (2015), pp. 1–36.
- [55] Tanja Sappok and Bernd Schmidt. “Psychische Gesundheit bei Personen mit intellektuellen Entwicklungsstörungen”. In: *PSYCH up2date* 15.03 (2021), pp. 199–214.
- [56] Andreas Säuberli, Sarah Ebling, and Martin Volk. “Benchmarking Data-driven Automatic Text Simplification for German”. In: *Proceedings of the 1st workshop on tools and resources to empower people with reading difficulties (READI)*. 2020, pp. 41–48.
- [57] Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. “Digital Comprehensibility Assessment of Simplified Texts among Persons with Intellectual Disabilities”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–11.
- [58] Carolina Scarton and Lucia Specia. “Learning Simplifications for Specific Target Audiences”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2018, pp. 712–718.

- [59] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. “Proximal Policy Optimization Algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).
- [60] Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. “Subjective Text Complexity Assessment for German”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022, pp. 707–714.
- [61] Kim Cheng Sheang and Horacio Saggion. “Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer”. In: *Proceedings of the 14th International Conference on Natural Language Generation*. 2021, pp. 341–352.
- [62] Zhengyan Shi, Adam X Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. “Instruction Tuning with Loss Over Instructions”. In: *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. 2024.
- [63] Advaith Siddharthan and Angrosh Mandya. “Hybrid Text Simplification Using Synchronous Dependency Grammars with Hand-written and Automatically Harvested Rules”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 2014, pp. 722–731.
- [64] “Significant regional inequalities in the prevalence of intellectual disability and trends from 1990 to 2019: a systematic analysis of GBD 2019”. In: *Epidemiology and psychiatric sciences* 31 (2022), e91.
- [65] Nicolas Spring, Annette Rios Gonzales, and Sarah Ebling. “Exploring German Multi-Level Text Simplification”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 2021, pp. 1339–1349.
- [66] Nicolas Spring, Marek Kostrzewa, David Fröhlich, Annette Rios, Dominik Pfütze, Alessia Battisti, and Sarah Ebling. “Analyzing Sentence Alignment for Automatic Simplification of German Texts”. In: *Emerging Fields in Easy Language and Accessible Communication Research*. Springer, 2023, pp. 339–369.
- [67] Nicolas Spring, Marek Kostrzewa, Annette Rios, and Sarah Ebling. “Ensembling and Score-Based Filtering in Sentence Alignment for Automatic Simplification of German Texts”. In: *International Conference on Human-Computer Interaction*. 2022, pp. 137–149.
- [68] Regina Stodden, Omar Momen, and Laura Kallmeyer. “DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 16441–16463.
- [69] Elior Sulem, Omri Abend, and Ari Rappoport. “BLEU is Not Suitable for the Evaluation of Text Simplification”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 738–744. DOI: [10.18653/v1/D18-1081](https://doi.org/10.18653/v1/D18-1081). URL: <https://aclanthology.org/D18-1081/>.

- [70] Julia Suter, Sarah Ebling, and Martin Volk. “Rule-based Automatic Text Simplification for German”. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. 2016, pp. 279–287.
- [71] Suha S Al-Thanyyan and Aqil M Azmi. “Automated Text Simplification: A Survey”. In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–36.
- [72] Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. “A New Aligned Simple German Corpus”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 11393–11412.
- [73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. “LLaMA: Open and Efficient Foundation Language Models”. In: *arXiv preprint arXiv:2302.13971* (2023).
- [74] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. *TRL: Transformer Reinforcement Learning*. Version 0.16. 2023. URL: <https://github.com/huggingface/trl>.
- [75] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. “A Survey of Human-in-the-loop for Machine Learning”. In: *Future Generation Computer Systems* 135 (2022), pp. 364–381.
- [76] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. “Optimizing Statistical Machine Translation for Text Simplification”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 401–415.
- [77] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 483–498.
- [78] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. “Including Signed Languages in Natural Language Processing”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 7347–7360. DOI: [10.18653/v1/2021.acl-long.570](https://doi.org/10.18653/v1/2021.acl-long.570). URL: <https://aclanthology.org/2021.acl-long.570/>.
- [79] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. “BERTScore: Evaluating Text Generation with Bert”. In: *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*. 2019.
- [80] Xingxing Zhang and Mirella Lapata. “Sentence Simplification with Deep Reinforcement Learning”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 584–594.



- [81] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. “Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs”. In: *Proceedings of the 15th International Conference on Learning Representations (ICLR 2025)*. 2025.
- [82] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. “LIMA: Less is More for Alignment”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023, pp. 55006–55021.

## **Appendix A**

### **SFT Grid Search for Parameter Mix**

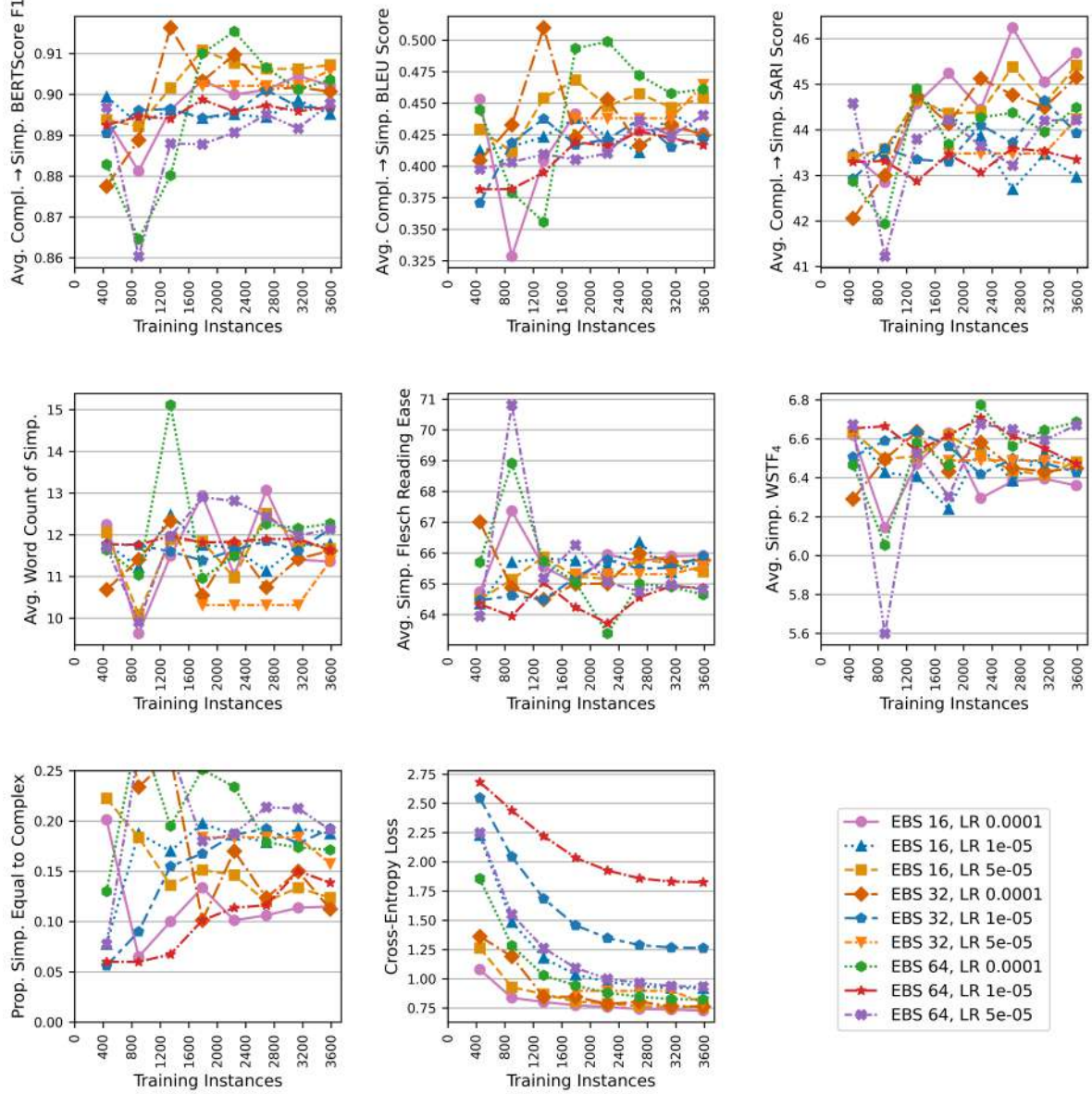


Figure A.1: SFT parameter mix grid search results for DiscoLeo-Llama-3-8B-Instruct. In the legend, EBS stands for Effective Batch Size, which equals the actual batch size of 16 multiplied by the gradient accumulation step size. We select gradient accumulation step size of 1 and learning rate of  $1e-4$  for the optimal parameter mix primarily based on SARI and Wiener Sachtextformel. We used full-prompt loss tuning during the grid search.

## **Appendix B**

### **ATS Pair Creation Filters**

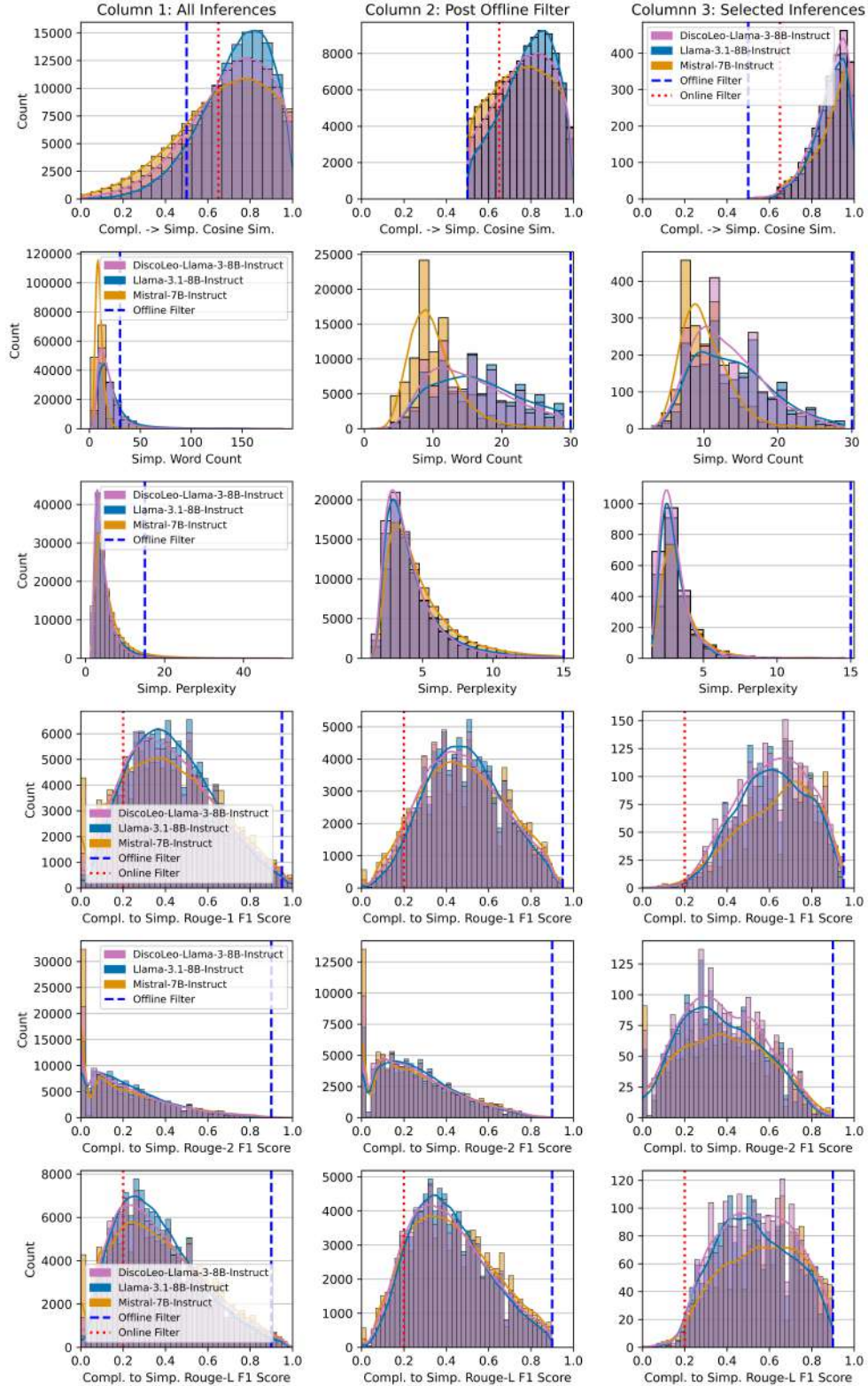


Figure B.1: Simplification-level distributions. To ensure we were not introducing bias into our ATS pairs, we verified that our selected simplifications (column 3) either left existing variable distributions unchanged or improved upon them by reducing skew.

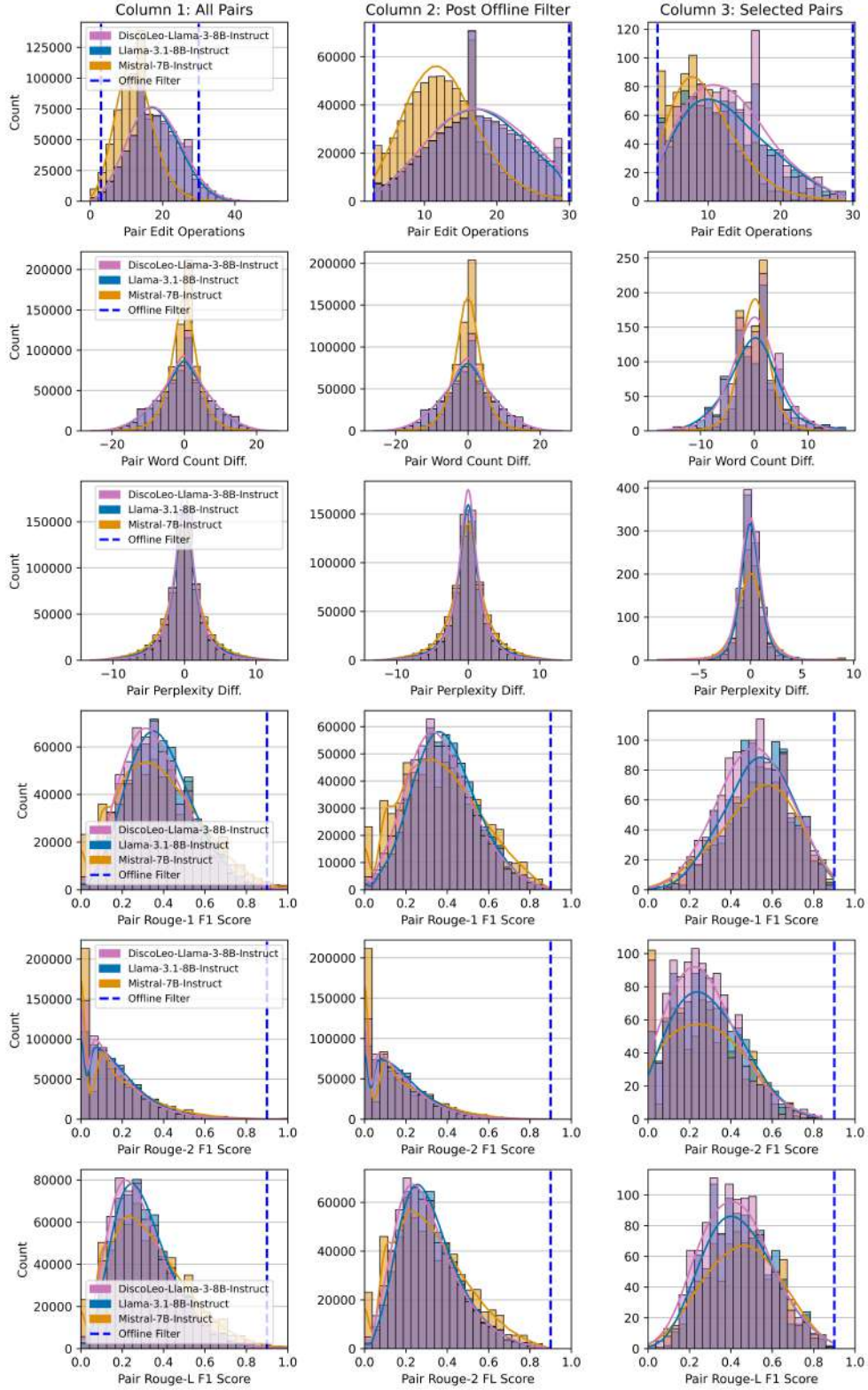


Figure B.2: Simplification-pair-level distributions. To ensure we were not introducing bias into our ATS pairs, we verified that our selected pairs (column 3) either left existing variable distributions unchanged or improved upon them by reducing skew.



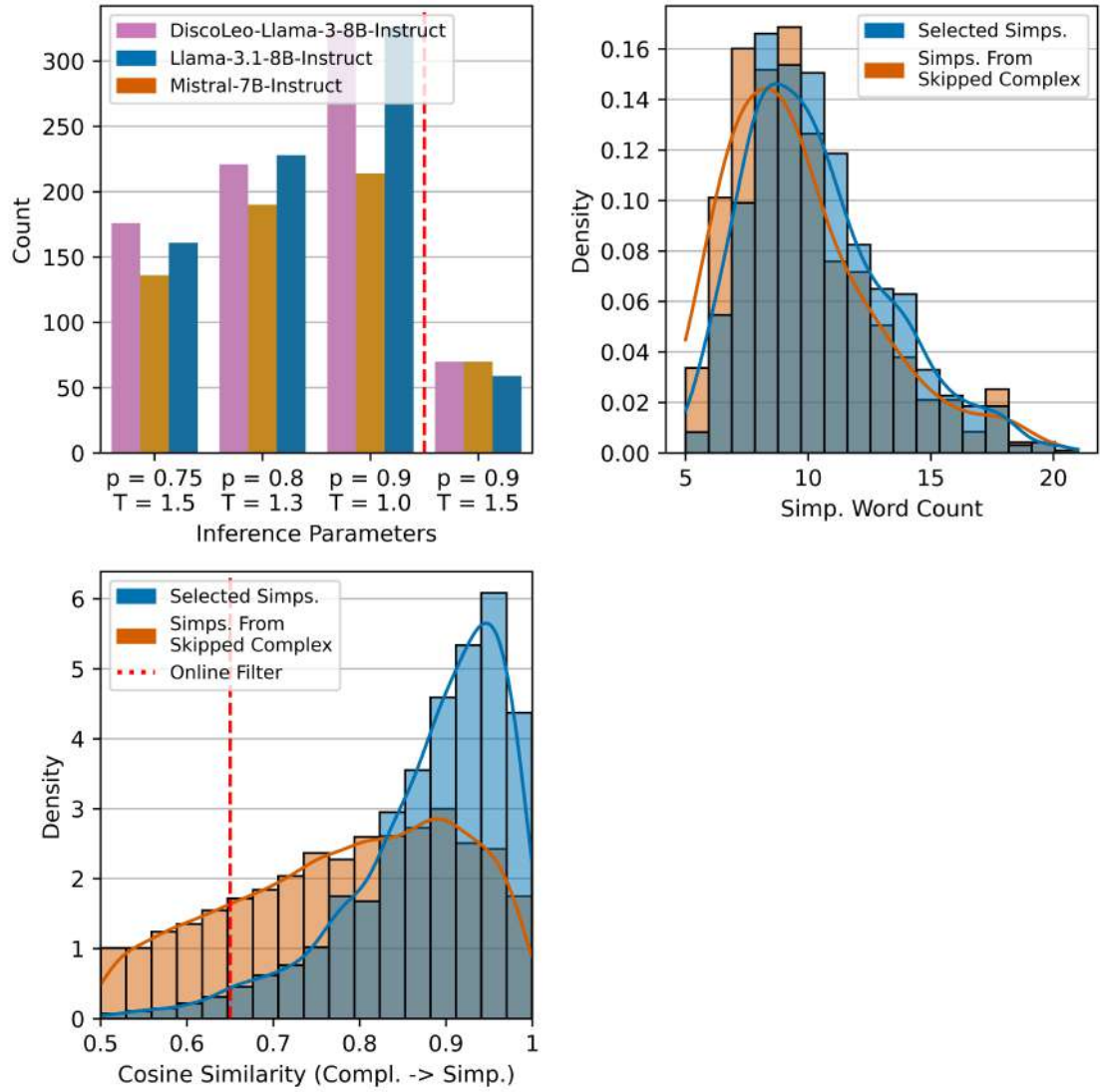


Figure B.3: Statistics for the first 1,030 created pairs. Inferences with temperature = 1.5 and  $p = 0.9$  during generation or with complex-simple cosine similarity less than 0.65 were excluded from future pair creation. No filtering occurred based on word count.

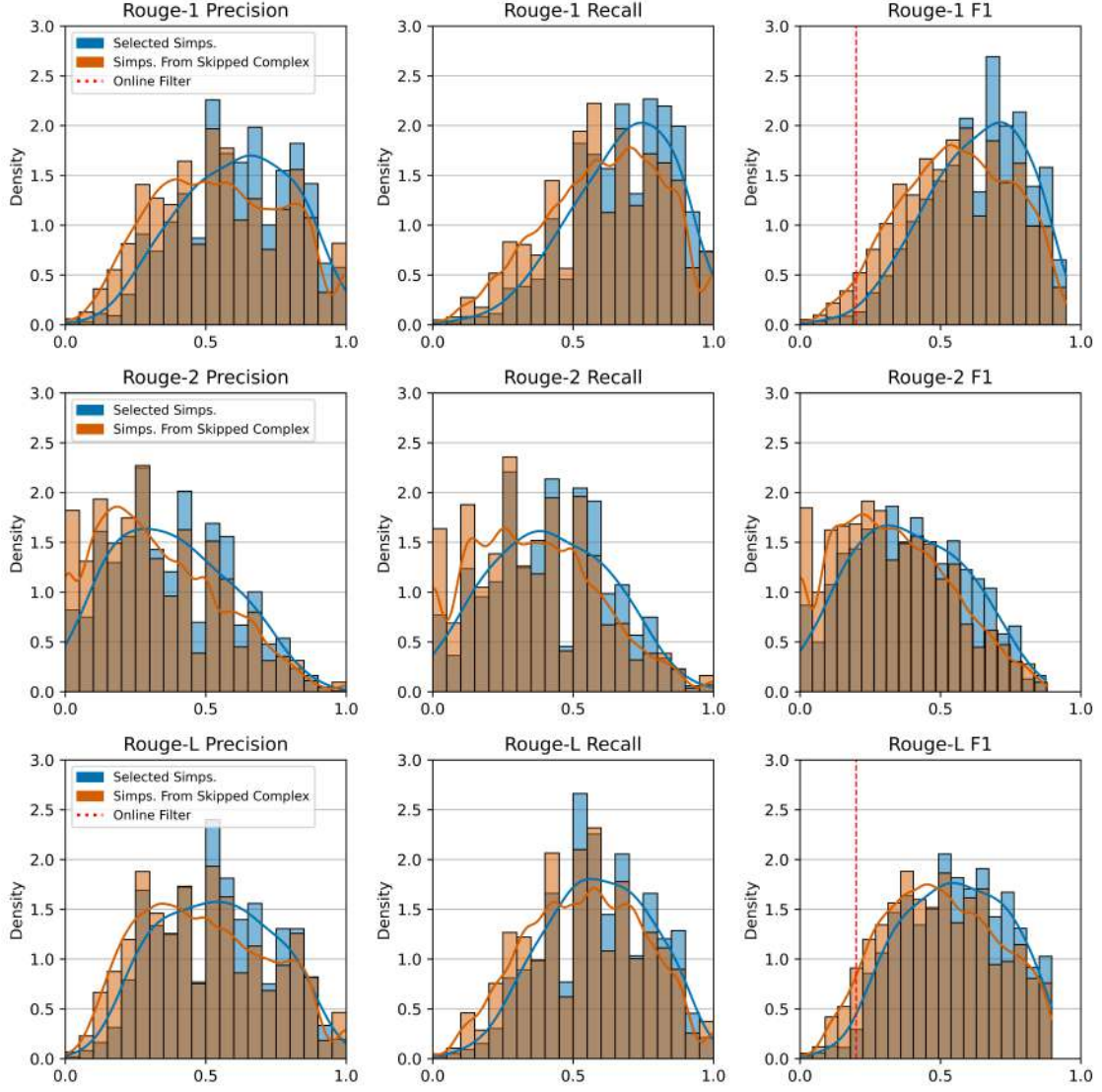


Figure B.4: Statistics for the first 1,030 created pairs. Inferences with complex-simple ROUGE-1 F1 or ROUGE-L F1 less than 0.2 were excluded from future pair creation.



## **Appendix C**

### **Annotation Session Documents**



**Universität  
Zürich** UZH

INSTITUT FÜR COMPUTERLINGUISTIK  
SPRACHE, TECHNOLOGIE UND BARRIEREFREIHEIT  
ANDREASSTRASSE 15, 8050 ZÜRICH, SCHWEIZ

## **Informations-Blatt für Teilnehmerinnen und Teilnehmer**

Hier bekommen Sie Informationen zu der Studie, an der Sie teilnehmen können.

Die Studie heisst:

*Welcher vereinfachte Text gefällt Ihnen besser?*

Die Studie macht Sarah Ebling.

Sie ist Professorin an der Universität Zürich in der Schweiz.

Zusammen mit capito arbeitet sie im Projekt

*Inclusive Information und Communication Technologies.*

Das ist englisch und bedeutet

*Informations- und Kommunikations-Technologien für alle.*

Sarah und ihr Team möchten wissen,

welche vereinfachten Texte Ihnen besser gefallen

und wie verständlich Texte sind.

Ihre Meinung ist uns wichtig.

Ihre Meinung hilft uns,

Texte noch verständlicher zu machen.

Dazu führen Sarah und ihr Team eine Studie durch.

Die Studie ist in 10 Sitzungen aufgeteilt.

Jede Sitzung dauert 1 Stunde.

Die Sitzungen finden über mehr als 8 Wochen verteilt statt.

Vor dem Studien-Anfang stellen wir Ihnen Fragen.

Zum Beispiel fragen wir Sie:

- Wie alt sind Sie?
- Welches Geschlecht haben Sie?
- Was ist Ihre Muttersprache?

Sie machen den Fragebogen nur einmal.

Sarah organisiert die Sitzungen so:  
Sie sehen zwei vereinfachte Texte,  
die von einem Computer geschrieben sind.  
Lesen Sie beide Texte durch.

Danach wählen Sie den Text, der Ihnen besser gefällt.  
Klicken Sie auf den entsprechenden Knopf.

Anschliessend sehen Sie 4 Fragen.  
Beantworten Sie die Fragen, indem Sie auf die passenden Emojis klicken.

Wenn Sie die Fragen beantwortet haben,  
machen Sie beim nächsten Text weiter.  
Sie lesen die Texte in Ihrem Tempo.  
Jede Sitzung ist gleich.

**Sarah und ihr Team versprechen Ihnen,  
dass sie und ihr Team Sie weder auf Video noch auf Ton aufnehmen.**

Ihre Freiwilligkeit ist Sarah sehr wichtig.  
Sie können sich selbst entscheiden:

- Möchte ich bei der Studie mitmachen?
- Oder möchte ich bei der Studie **nicht** mitmachen?

Auch nach dem Start können Sie Sarah jederzeit sagen:

- Ich möchte doch **nicht** mitmachen.
- Ich höre jetzt wieder auf.

Wenn Sie bei der Studie mitmachen möchten, melden Sie sich bei Ihrer capito-Person.

Wenn Sie mitmachen, bekommen Sie Geld.

So, wie wenn Sie bei einer Prüfgruppe mitmachen.

Wenn Sie Fragen zur Studie haben,  
melden Sie sich bitte bei der Studien-Leiterin:

Prof. Dr. Sarah Ebling

Institut für Computerlinguistik, Universität Zürich  
Andreasstrasse 15, CH-8050 Zürich  
E-Mail: [ebling@cl.uzh.ch](mailto:ebling@cl.uzh.ch)



Universität  
Zürich<sup>UZH</sup>

INSTITUT FÜR COMPUTERLINGUISTIK  
SPRACHE, TECHNOLOGIE UND BARRIEREFREIHEIT  
ANDREASSTRASSE 15, 8050 ZÜRICH, SCHWEIZ

## Einwilligungs-Erklärung

Unsere Studie heisst:

*Welcher vereinfachte Text gefällt Ihnen besser?*

Bevor Sie mitmachen,

müssen Sie diese Einwilligungs-Erklärung unterschreiben.

Wenn Sie unterschreiben, heisst das:

**Sie sind mit allem einverstanden,  
was in der Einwilligungs-Erklärung steht.**

Die Studie macht Sarah Ebling.

Sie ist Professorin an der Universität Zürich in der Schweiz.

Sie arbeitet mit capito im Projekt

*Inclusive Information and Communication Technologies* zusammen.

Das ist englisch und bedeutet

*Informations- und Kommunikations- Technologien für alle.*

Sarah und ihr Team speichern

während der Studie die Daten über Sie.

Zum Beispiel:

- Ihren Namen
- Ihr Alter
- Ihre Antworten

---

Ich bin damit einverstanden,

dass Sarah und ihr Team meine Daten sammeln dürfen.

Meine Antworten und mein Name werden nie gemeinsam gespeichert.

Meine Antworten werden mit einem Code gespeichert.

Nur Sarah und ihr Team wissen,

welcher Benutzer-Name zu welchem echten Namen gehört.

Wenn ich wissen möchte,  
welche Daten über mich gespeichert sind,  
kann ich Sarah jederzeit fragen.

Meine Daten werden auf Computern  
von der Universität Zürich in der Schweiz gespeichert und bearbeitet.  
Sarah passt gut auf meine Daten auf.  
Zum Beispiel schaut Sarah, dass

- **kein** anderer Mensch die Daten lesen kann.
- **kein** anderer Mensch die Daten stehlen kann.

Nur bestimmte Mitarbeiterinnen und Mitarbeiter aus  
Sarahs Team dürfen mein Daten lesen.

Sie dürfen über meine Daten mit **keinem** anderen Menschen reden.

Im Februar 2026 hört das Projekt auf.  
Dann wird die Liste gelöscht,  
auf der steht, welcher echte Name zu welchem Benutzer-Namen gehört.  
Dann sind meine Daten ganz anonym.

Die Daten werden für mindestens 10 Jahre gespeichert.  
Ich bin damit einverstanden,  
dass meine Daten in dieser Zeit für die Forschung verwendet werden.  
Sie dürfen für nichts anderes verwendet werden.

Wir wollen Ihre Daten nur **anonym** veröffentlichen.  
Das heisst, **niemand** kann Ihren Namen, Ihr Alter  
oder andere persönliche Informationen sehen.

Die Daten dürfen nur für die Forschung verwendet werden.  
Das heisst, **niemand** darf Geld mit den Daten machen.

Ihre Daten speichern wir nur auf europäischen Datenbanken

wie SwissUbase oder Zenodo, nirgendwo anders.

Wenn Ihre Daten veröffentlicht sind,  
können wir sie **nicht** mehr löschen.

Wenn Sie unterschreiben,  
heisst es, dass Sie damit einverstanden sind,  
Ihre Daten **anonym** zu veröffentlichen.

---

Ich hatte genügend Zeit.  
Ich konnte mir überlegen:  
Möchte ich bei der Studie mitmachen?  
Das ist freiwillig.  
Ich weiss:  
Ich kann meine Meinung immer ändern.  
Ich kann immer sagen:  
Ich möchte doch **nicht** mehr mitmachen.  
Das ist nicht schlimm.  
Das ist **kein** Nachteil für mich.  
Ich muss das nicht erklären.

Wenn ich nicht mehr mitmachen möchte,  
melde ich es Sarah.  
Wenn ich das Projekt dann noch nicht vorbei ist,  
löscht Sarah meine Daten.  
Wenn das Projekt dann schon vorbei ist,  
sind meine Daten schon anonym.  
Dann kann Sarah sie nicht mehr löschen.

Ich habe ein Informations-Blatt bekommen.  
Das Blatt gehört auch zu dieser Einwilligungs-Erklärung.

Jemand hat mir **genau** erklärt,  
wie die Studie abläuft.  
Ich konnte Fragen stellen.  
Ich habe alles verstanden.  
Ich habe **keine** Fragen mehr zur Studie.

Ich bekomme eine Kopie von dieser Einwilligungs-Erklärung.

Ich möchte gerne bei der Studie mitmachen: ☐ Ja ☐ Nein

---

Ort, Datum

---

Name des Teilnehmers oder der Teilnehmerin

---

Unterschrift des Teilnehmers oder der Teilnehmerin

---

Ort, Datum

---

Unterschrift der Studienleiterin

Studienleiterin:

Prof. Dr. Sarah Ebling

Universität Zürich

Institut für Computerlinguistik

Andreasstrasse 15, CH-8050 Zürich

E-Mail: ebling@c1.uzh.ch

Wenn ich mich über die Studie beschweren möchte,  
kann ich das hier tun:

Prof. Dr. Christopher Hopwood

Präsident der Etikkommission



Psychologisches Institut  
Binzmühlestrasse 14, P.O. Box 34  
CH-8050 Zürich  
[chair.ethics.committee@phil.uzh.ch](mailto:chair.ethics.committee@phil.uzh.ch)



**Universität  
Zürich** UZH

INSTITUT FÜR COMPUTERLINGUISTIK  
SPRACHE, TECHNOLOGIE UND BARRIEREFREIHEIT  
ANDREASSTRASSE 15, 8050 ZÜRICH, SCHWEIZ

## **Schriftliche Instruktionen für Teilnehmende (provisorisch, noch in Leichte Sprache zu übertragen)**

Herzlich willkommen bei unserer Studie!

Bitte gehen Sie in Ihrem Internet-Browser auf diesen Link:

<https://pub.cl.uzh.ch/projects/dpo4ats/>

Wir empfehlen den Chrome-Browser für die Teilnahme an der Studie.

1. Bitte geben Sie Ihre User-ID ein. Ihre User-ID ist: **ta01**

Teilen Sie Ihre User-ID **nicht** mit anderen Teilnehmenden.

Danach klicken Sie auf den blauen Button "Einloggen".

Bitte geben Sie Ihre User-ID ein ...

EINLOGGEN

2. Sie werden ein paar Informationen sehen.

Unten sehen Sie auch zwei Texte, die von einem Computer geschrieben sind.

Jeder Text hat am Ende einen grünen Button.

Herzlich Willkommen bei unserer Studie!

Wir verwenden Ihre Antworten,  
um die KI zu trainieren,  
damit die KI bessere Texte schreiben kann.

Klicken Sie auf den ABSCHICKEN Button,  
nachdem Sie Ihre Antworten gegeben haben.

Sie können jeder Zeit Ihre Antworten abschicken  
Machen Sie eine Pause, wenn Sie müde sind.

Sarah Ebling bedankt sich für Ihre Teilnahme!

ZURÜCKWEITER

Text 1:

Alle Praktika und Studienaufenthalte, die du bisher mit den Programmen Erasmus+, Lebenslanges Lernen (LLP-Erasmus) oder Erasmus Mundus gemacht hast, müssen bei der Planung deines neuen Aufenthalts gezählt werden.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Alle Praktika und Studien, die du schon mit Erasmus+, Lebenslanges Lernen oder Erasmus Mundus gemacht hast, zählen zur Dauer deines neuen Aufenthalts dazu.

DIESEN TEXT VERSTEHE ICH BESSER

3. Lesen Sie beide Texte durch.

Wählen Sie den Text, der Ihnen besser gefällt und klarer ist.

Wenn Sie wählen, klicken Sie auf den grünen Button im Text.

Wenn Sie auf den grünen Button klicken, bekommt der Text einen grünen Hintergrund.

The screenshot shows a web form with two text boxes, 'Text 1' and 'Text 2', each containing a paragraph of text and a green button labeled 'DIESEN TEXT VERSTEHE ICH BESSER'. Above the text boxes is a blue button labeled 'ABSCHICKEN'. At the top left is a grey button labeled 'ZURÜCK' and at the top right is a blue button labeled 'WEITER'.

Text 1:

Alle Praktika und Studienaufenthalte, die du bisher mit den Programmen Erasmus+, Lebenslanges Lernen (LLP-Erasmus) oder Erasmus Mundus gemacht hast, müssen bei der Planung deines neuen Aufenthalts gezählt werden.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Alle Praktika und Studien, die du schon mit Erasmus+, Lebenslanges Lernen oder Erasmus Mundus gemacht hast, zählen zur Dauer deines neuen Aufenthalts dazu.

DIESEN TEXT VERSTEHE ICH BESSER

ABSCHICKEN

ZURÜCK

WEITER

Danach sehen Sie den blauen Button "Abschicken". Das ist normal.

Sie können jeder Zeit Ihre Wahl abschicken.

4. Haben Sie Ihre Wahl abgeschickt?

Dann klicken Sie auf den blauen Button "Weiter"

und machen Sie die nächsten Texte.

Wenn Sie Ihre Antworten ändern wollen,

klicken Sie auf den Button "Zurück",

so kommen Sie zurück zu den letzten Texten.

The screenshot shows the same web form as before, but with the 'WEITER' button at the top right highlighted with a red oval. The 'ABSCHICKEN' button is still present above the text boxes.

Text 1:

Alle Praktika und Studienaufenthalte, die du bisher mit den Programmen Erasmus+, Lebenslanges Lernen (LLP-Erasmus) oder Erasmus Mundus gemacht hast, müssen bei der Planung deines neuen Aufenthalts gezählt werden.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Alle Praktika und Studien, die du schon mit Erasmus+, Lebenslanges Lernen oder Erasmus Mundus gemacht hast, zählen zur Dauer deines neuen Aufenthalts dazu.

DIESEN TEXT VERSTEHE ICH BESSER

ABSCHICKEN

ZURÜCK

WEITER

5. Wenn Sie nicht weiter machen,

**klicken Sie unbedingt noch mal auf den blauen Button “Abschicken”.**

So haben wir Ihre aktuelle Antworten.



The form interface includes three blue buttons: 'ZURÜCK' on the left, 'ABSCHICKEN' in the center (highlighted with a red dashed border), and 'WEITER' on the right. Below these buttons are two text input fields: 'Text 1:' with a green background and 'Text 2:' with a dark grey background.

Wenn Sie diese Nachricht sehen,

sind Ihre Antworten bei uns gut angekommen.



Bitte nehmen Sie sich Zeit.

Wir erwarten nicht, dass Sie alle Texte in einer Sitzung sehen.

Sarah und ihr Team bedanken sich herzlich für Ihre Teilnahme!



**Universität  
Zürich** UZH

INSTITUT FÜR COMPUTERLINGUISTIK  
SPRACHE, TECHNOLOGIE UND BARRIEREFREIHEIT  
ANDREASSTRASSE 15, 8050 ZÜRICH, SCHWEIZ

## Fragebogen

Wie ist Ihr Name? \_\_\_\_\_

Wie alt sind Sie? \_\_\_\_\_ Jahre

Geschlecht: ☐ männlich ☐ weiblich ☐ divers

Was ist Ihre Muttersprache? \_\_\_\_\_

Sprechen Sie weitere Sprachen? Welche? \_\_\_\_\_

Wie oft lesen Sie Texte (zum Beispiel in einer Zeitung,  
in einem Buch oder im Internet)?

- ☐ Jeden Tag
- ☐ Ein paar Mal pro Woche
- ☐ Ein paar Mal pro Monat
- ☐ Weniger

Welches Gerät nutzen Sie heute?

- ☐ Laptop
- ☐ PC
- ☐ Tablet

Wie oft lesen Sie Texte in Leichter Sprache?



- ☐ Jeden Tag
- ☐ Ein paar Mal pro Woche
- ☐ Ein paar Mal pro Monat
- ☐ Weniger

Wenn Sie in Leichter Sprache lesen: Wer macht die Leichte Sprache für Sie?

- ☐ capito
- ☐ Medien (zum Beispiel APA oder Tagesschau aus Deutschland)
- ☐ Künstliche Intelligenz/Computer (zum Beispiel: ChatGPT)
- ☐ Andere (bitte schreiben \_\_\_\_\_)

## Appendix D

# Web Application



Universität Zürich  
Department für Computerlinguistik

Zürcher Hochschule für Angewandte Wissenschaften  
Department für Angewandte Linguistik

Sie sind eingeloggt: ta02

Herzlich Willkommen bei unserer Studie!

Wir verwenden Ihre Antworten,  
um die KI zu trainieren,  
damit die KI bessere Texte schreiben kann.

Klicken Sie auf den ABSCHICKEN Button,  
nachdem Sie Ihre Antworten gegeben haben.

Sie können jeder Zeit Ihre Antworten abschicken  
Machen Sie eine Pause, wenn Sie müde sind.

Sarah Ebling bedankt sich für Ihre Teilnahme!

ZURÜCK

WEITER

Text 1:

Matteo Salvini ist der Innen-Minister von Italien  
und Chef der Lega.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Matteo Salvini ist Innen-Minister in Italien. Er ge-  
hört zur Partei Lega.

DIESEN TEXT VERSTEHE ICH BESSER

Figure D.1: Full view of annotation web application after sign-in.



Figure D.2: Partial view of annotation web application after selecting a preference.



Figure D.3: Partial view of annotation web application for annotators ea02 and ea04.



## **Appendix E**

### **Problematic Pairs**

Text 1:

Baby-Elefant nennt man den Mindest-Abstand. Er gibt es seit der Corona-Krise. Bei Baby-Elefant soll man die Entfernung zu anderen Menschen haben. Man soll 1,5 Meter von ihnen fern bleiben.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Der Begriff "Baby-Elefant" soll den Mindest-Abstand bezeichnen, den man seit der Corona-Krise in der Öffentlichkeit einhalten muss.

DIESEN TEXT VERSTEHE ICH BESSER

links: Grammatikfehler und untypische Formulierungen; rechts: kompliziertere Satzstruktur → schwer zu sagen, was besser ist

Text 1:

In vielen Orten haben die Flüsse über die Ufer getreten. Diese überfluteten ganze Wohn-Gebiete und Straßen.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Viele Flüsse traten über die Ufer und überschwemmten Städte und Wohn-Gebiete.

DIESEN TEXT VERSTEHE ICH BESSER

links: Grammatikfehler; rechts: evtl. störender Zeilenumbruch

Text 1:

Denn das BVT ist mit dem Internet verbunden.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Das Computer-Netzwerk des BVT ist mit dem Internet verbunden.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: evtl. störender Zeilenumbruch

Text 1:

Die Polizei denkt, dass es ein Terror-Anschlag war.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Die Polizei glaubt, dass es einen Terror-Anschlag war.

DIESEN TEXT VERSTEHE ICH BESSER

nur ein Wort anders, rechts: Grammatikfehler – da sucht man dann nicht nach Verständlichkeit aus, sondern nimmt eher links, weil's fast gleich ist und keinen Fehler hat

<p>Text 1:</p> <p>Der Fix-Kosten-Zuschuss ist eine Hilfe für die Ge- schäfte. Der Schutz-Schirm soll zusätzlich helfen, sagte Blümel. Der Schutz-Schirm bietet im Vorhin- ein Planung-Sicherheit.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Es solle sozusagen eine Hilfe im Nachhinein sein. Ein anderer Schutzschirm soll helfen, im Voraus zu wissen, wie viel Geld man erhält. Damit will Blümel die Menschen beschützen.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	---

links: evtl. störender Zeilenumbruch, Grammatikfehler

<p>Text 1:</p> <p>Die Politiker im Kongress kontrollieren den Präsidenten.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Im Kongress entscheiden die Politiker mit.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
--	--

anderer Inhalt

<p>Text 1:</p> <p>Leben in Österreich kostet immer mehr Geld.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>In Österreich ist das Wohnen jetzt teurer als noch vor einem Jahr.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	--

anderer Inhalt

<p>Text 1:</p> <p>Die kühle Woche hat die Sache nicht verzögert.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Auch diese relativ kalten Tage dieser Woche ha- ben da nicht geholfen.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
--	--

beides schwer (links: „verzögert“ = schweres Wort, „die Sache = unkonkret;  
recht: 2 Mal „diese“, untypische Formulierung)

Text 1:

Kurz sagte auch, dass viele Menschen nicht glauben würden, dass die Infektions-Zahlen so hoch sind.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Kurz sagte, dass viele Menschen das nicht glauben würden. Es würde nämlich viele Menschen geben, die behaupten, dass die Infektionszahlen nicht so hoch sind.

DIESEN TEXT VERSTEHE ICH BESSER

beides schwer (Konjunktiv, komplexe Sätze und Wörter)

Text 1:

Es gibt also noch keinen Kontakt zwischen den verfeindeten Parteien. Die Vereinten Nationen sagten dies.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

In den Vereinten Nationen sagte man zunächst, dass keine direkten Verhandlungen geplant sind.

DIESEN TEXT VERSTEHE ICH BESSER

beides schwer (komplexe Wörter)

Text 1:

Dann müssen die Angestellten vielleicht eine Schutz-Maske tragen.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Die Mitarbeiter müssen Schutz-Masken tragen.

DIESEN TEXT VERSTEHE ICH BESSER

andere Aussage

Text 1:

In Deutschland vermisste ein Mann sein Auto. Er suchte das Auto 3 Wochen lang.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Ein Mann in Deutschland verlor **seinem** Auto. Der Mann suchte nach seinem Auto. Er fand es erst 3 Wochen später wieder.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:

Das Gemälde wurde am Abend auf einer Auktion für Rekord-Preise versteigert. Das schrieb der Künstler Banksy. Er postete das auf seinem Instagram-Konto.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Heute Abend versteigerte der Künstler ein Bild. Es stellte einen neuen Rekord auf.

DIESEN TEXT VERSTEHE ICH BESSER

links: mehr Info

Text 1:

Trump legte sich zudem mit dem Londoner Bürgermeister Sadiq Khan an.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Trump machte auch einen Streit mit dem Bürgermeister von London, Sadiq Khan.

DIESEN TEXT VERSTEHE ICH BESSER

schwere Entscheidung: links: „sich mit XY anlegen“ = schwer, rechts: „Streit machen“ ungebräuchlich; beide Seiten: Zeilenumbrüche

Text 1:

LASK gewinnt überraschend in der Champions League.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Das Team LASK wird an der Champions League spielen. Das war nicht vorher abzusehen.

DIESEN TEXT VERSTEHE ICH BESSER

andere Aussage; rechts: Grammatikfehler

Text 1:

Eisbären sind bei der Geburt so groß wie ein kleines Meerschweinchen und wiegen ungefähr so viel wie ein Meerschweinchen.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Junge Eisbären kommen etwa ein halbes Kilo auf die Welt und sind dann so groß wie ein Meerschweinchen.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:

In Österreich hat sich in den letzten 24 Stunden 272 Menschen mit dem Corona-Virus angesteckt.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

In den letzten 24 Stunden gab es in Österreich 272 Neu-Infektionen mit dem Corona-Virus.

DIESEN TEXT VERSTEHE ICH BESSER

schwer zu entscheiden: links: Grammatikfehler; rechts: schweres Wort (Neu-Infektionen)

Text 1:

Die Busch-Brände in Australien dauern schon seit rund 2 Monaten an.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Seit 2 Monaten brennen die Busch-Brände in Australien.

DIESEN TEXT VERSTEHE ICH BESSER

schwer zu entscheiden: links: „andauern“ = schwer, rechts: untypische Formulierung „die Brände brennen“

Text 1:

Der Burgenländische Landtag hat an Montag gewählt. Er wählte Hans Peter Doskozil erneut zum Landeshauptmann. Doskozil gehört der SPÖ an.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Der Burgenländische Landtag hat Hans Peter Doskozil wiedergewählt. Hans Peter Doskozil ist Landeshauptmann von Burgenland. Er stammt von der Partei SPÖ.

DIESEN TEXT VERSTEHE ICH BESSER

schwer zu entscheiden: Grammatikfehler und Zeilenumbrüche in beiden Texten; rechts: untypische Formulierung („stammt von der Partei“)

Text 1:

Er sagt, dass man die Sport-Stätten schon jetzt wieder öffnen kann. Aber man muss das klug machen.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Wenn man diese Regeln befolgt, kann man Sportstätten bald wieder öffnen.

DIESEN TEXT VERSTEHE ICH BESSER

anderer Inhalt

Text 1:  
Diese Partei wird das Team von Frau von der Leyen unterstützen.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
In dem Licht wird ihre Fraktion Team von Ursula von der Leyen unterstützen.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: „das“ fehlt, „in dem Licht“ = untypisch

Text 1:  
In Stockerau in Niederösterreich haben vor 2 Monaten Geldautomaten 50 statt 20 Euro ausgegeben.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
In Stockerau in Niederösterreich hat ein Bankomat vor etwa 2 Monaten 50- statt 20-Euro-Scheine ausgegeben.  
DIESEN TEXT VERSTEHE ICH BESSER

Zeilenumbruch

Text 1:  
Am 3. November 2020 wird in den USA ein neuer Präsident gewählt. Dazu treten die aktuelle Amtsinhaberin Trump und der Demokrat Joe Biden gegeneinander an.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
In den USA wird bald ein neuer Präsident gewählt. Das ist im November 2020.  
DIESEN TEXT VERSTEHE ICH BESSER

links: mehr Infos, „die Amtsinhaberin Trump“

Text 1:  
Douglas konnte den Oscar bisher 3 Mal gewinnen. Er war 3 Mal nominiert.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
Schon 3 Mal wurde der Star Douglas von der Oscar-Academy nominiert. Der Begriff „nominiert“ heißt in Deutsch: Ausgezeichnet.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: Fehler, denn nominiert heißt nicht ausgezeichnet; links: nominiert und gewinnen ist nicht das Gleiche

Text 1:

Sollte diese Entwicklung weitergehen, könnte es im intensiv-medizinischen Bereich kritisch werden.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Wenn das so weiter geht, dann könnte es im intensiv-medizinischen Bereich eng werden.

DIESEN TEXT VERSTEHE ICH BESSER

beides schwer wegen „intensiv-medizinischer Bereich“, rechts: Zeilenumbruch

Text 1:

Der Innenminister von Frankreich hat die Tat verurteilt. Er sagte, dass er den muslimischen Gemeinschaften Unterstützung und Solidarität anbietet.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Der französische Innen-Minister hat die Tat verurteilt. Er sagte, dass die muslimische Gemeinschaft Solidarität haben wird.

DIESEN TEXT VERSTEHE ICH BESSER

2. Satz bei beiden schwer und nicht idiomatisch

Text 1:

Thomas Goiginger spielte zum ersten Mal in dieser Saison für den LASK in der Startelf.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Beim LASK spielte Goiginger beim 1. Spiel in dieser Saison in der Startelf.

DIESEN TEXT VERSTEHE ICH BESSER

unterschiedliche Aussagen

Text 1:

Daraufhin zogen die Rettungskräfte Leichen aus dem Fluss. Sie übergaben die Leichen den Angehörigen.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Augenzeugen sagten, dass die Retter Körper aus dem Fluss fischten. Sie übergaben die Todesopfer dann an deren Angehörige.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: Körper = nicht Leiche; „fischen“ passt nicht

Text 1:

Teilweise fehlte es auch an Bewegung. Foda wirkte unzufrieden.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Sie hatten keinen Bewegung. Man sah Foda an, dass er unglücklich war.

DIESEN TEXT VERSTEHE ICH BESSER

links: „fehlte es an“ = schwer; rechts: Grammatikfehler



Text 1:

In großen Gruppen zogen sie durch die Straßen.  
Dort plünderten sie Geschäfte.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

In großen Gruppen zogen sie durch die Stadt,  
plünderten Geschäfte und Läden. Dabei machten  
sie auch viel Chaos.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: mehr Infos

Text 1:

Viele Frauen in Österreich fühlen sich gestresst.  
Grund dafür sind die Home-Office-Arbeit und das  
Home-Schooling für die Kinder. Laut AK ist das  
die häufigste Antwort.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Home-Office und Home-Schooling stressten  
Frauen am meisten. Das sagt der AK

DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:

Er wollte, dass die Diskussion nach Regeln läuft  
und dass jeder seine Meinung ausdrücken darf.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Er möchte, dass die Diskussion nach Regeln er-  
folgt. So soll jeder seine Meinung sagen können.

DIESEN TEXT VERSTEHE ICH BESSER

beides schwere Formulierungen (links: „Meinung ausdrücken“, rechts: „nach Regeln erfolgt“)

Text 1:

Red Bull Salzburg hatte am Dienstag sein erstes  
Spiel in der Champions-League. Das war sehr  
magisch.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Bei Red Bull Salzburg ist eine magische Nacht in  
der Fußball-Champions-League zu Ende  
gegangen.

DIESEN TEXT VERSTEHE ICH BESSER

unterschiedliche Infos

<p>Text 1:</p> <p>So sah sich Paenda mit einigen Favoriten von der heurigen ESC-Königin konfrontiert.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Paenda musste dann auch gegen viele der Favoriten auf die heurige ESC-Krone singen.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	---

beides untypische Formulierungen (links: „mit Favoriten konfrontiert sehen“ im Sinne von gegeneinander antreten, was sind „Favoriten der heurigen ESC-Königin“?; rechts: „auf die ESC-Krone singen“)

<p>Text 1:</p> <p>Ich bitte alle, die Regeln zu beachten.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Ich bitte auch alle, die Abstands-Regeln einzuhalten.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	---

Abstand kommt links nicht vor – andere Infos

<p>Text 1:</p> <p>Man wählt sie in den Nationalrat.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Bei den Nationalrats-Wahlen werden die Abgeordneten gewählt.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	--

andere Aussage

<p>Text 1:</p> <p>Das Nein zum Freihandels-Abkommen hat sich auch der Handels-Verband gefreut. Er nannte das Nein für den Handel wichtig.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Daher ist auch der Handels-Verband zufrieden. Der Handels-Verband hat das nein zum Abkommen als Sieg gewertet.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	--

links: Grammatikfehler, rechts: schwere Formulierung

Text 1:

Damit sich die Wege der Besucher nicht ständig kreuzen, gibt es auch einen Rund-Weg. Damit wollen sie die Gesundheit der Menschen schützen.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Daher wird ein Rund-Weg angeboten. Damit vermeidet man, dass sich die Wege der Besucher ständig kreuzen.

DIESEN TEXT VERSTEHE ICH BESSER

links: mehr Info

Text 1:

Schwere Unwetter und Überschwemmungen haben am Wochenende in Süd-Frankreich und in Nord-Italien für viele Schäden gesorgt.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Am Wochenende haben in Italien und Frankreich schwere Unwetter und Überschwemmungen für viele Probleme gesorgt.

DIESEN TEXT VERSTEHE ICH BESSER

beides enthält schwere Formulierungen

Text 1:

Tausende Menschen ehren Niki Lauda.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Tausende Menschen haben Niki Lauda **ehrt**. Sie wollten zeigen, dass sie Niki Lauda schätzten.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:

Mancherorts wurden die Dächer von den Häusern gerissen. Straßen und Wohnhäuser wurden teilweise überschwemmt.

DIESEN TEXT VERSTEHE ICH BESSER

Text 2:

Straßen und Wohnhäuser wurden vom Wasser und Schlammmassen beschädigt. An vielen Stellen wurden Dächer vom Wasser weggewaschen.

DIESEN TEXT VERSTEHE ICH BESSER

rechts: untypische Formulierung „Dächer weggewaschen“

<p>Text 1:</p> <p>Die beiden Vermissten wurden schließlich gefunden. Die Retter konnten sie mit LVS-Geräten ausgraben.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>	<p>Text 2:</p> <p>Schließlich fanden sie die beiden Vermissten mit einem Suchgerät, das LVS-Gerät heißt. Sie gruben die beiden Vermissten mit einer Schaufel aus.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>
---	--

andere Info – links: ausgraben mit LVS-Geräten, rechts: ausgraben mit Schaufel

<p>Text 1:</p> <p>Jetzt sind die 6 Baby-Präriehunde schon größer. Sie unternehmen schon längere Ausflüge.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>	<p>Text 2:</p> <p>Die 6 jungen Präriehunde haben inzwischen auch regelmäßig lange Ausflüge. Sie erkunden schon selbstständig die Gehege.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>
--	---

rechts: untypische Formulierung „Ausflüge haben“

<p>Text 1:</p> <p>Die PlayStation 5 wird es in 2 Varianten geben.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>	<p>Text 2:</p> <p>Die PlayStation 5 wird es in 2 Versionen geben. Es wird ein normales Gerät und eine Version mit dem Namen PlayStation 5-Digital Edition geben.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>
--	---

erster Satz quasi gleich; rechts: mehr Info

<p>Text 1:</p> <p>Gleichzeitig soll in den Umbau in eine digitalere und klimafreundlichere Wirtschaft investiert werden.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>	<p>Text 2:</p> <p>Dafür soll auch in die Wirtschaft investiert werden. Sie soll digitaler und auch freundlicher zu der Umwelt werden.</p> <p><a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a></p>
---	--

schwer zu entscheiden: links: schwere Wörter, rechts: untypische Formulierung „freundlicher zu der Umwelt“

Text 1:  
Ausgenommen sind Osttirol und Kinder bis 10 Jahre. Diese brauchen keinen Test.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
Ausgenommen sind dabei die Gebiete Osttirol und Kindergarten- und Schüler bis zehn Jahre. Diese müssen keinen Test vorweisen.  
DIESEN TEXT VERSTEHE ICH BESSER

links: untypische Formulierung „Osttirol und Kinder“ – als wäre das die gleiche Kategorie; rechts: „Gebiete“ im Plural passt nicht, Tippfehler „Kindergarten-“,

Text 1:  
Man muss den Wunsch nach einem Papiamont dem Arbeitgeber vorher mitteilen. Dafür gibt es eine bestimmte Frist.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
Für einen Papi-Monat gibt es jedoch Vorschriften. Arbeitgeber und Arbeitnehmer müssen ihn mitteilen. Das darf frühestens drei Monate im Voraus.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:  
Im Freien darf eine Gruppe von bis zu 12 Menschen zusammen sein.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
Im Freien können bis zu 12 Personen miteinander treffen.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:  
In den Jahren davor ist dieser Anteil immer höher geworden. Im Vorjahr lag er schon bei 56 Prozent.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
Dieser Anteil ist in den letzten Jahren immer höher geworden. Im Vorjahr war es schon 56 Prozent.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: Grammatikfehler

Text 1:  
Der Bart von den drei kleinen Winzlingen wächst auch gut, sagt der Tiergartendirektor.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
Auch bei den 3 Winzlingen sieht man den Bartansatz schon gut. Das sagt der Tiergartendirektor.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: besonders irritierender Zeilenumbruch

Text 1:	Text 2:
Greta Thunberg ist ein Mädchen aus Schweden. Sie setzt sich für den Klimaschutz ein. Am Mittwoch hat sie Papst Franziskus getroffen.	Die schwedische Umwelt-Aktivistin Greta Thunberg hat Papst Franziskus getroffen. Das Treffen fand am Mittwoch nach der General-Audienz auf dem Petersplatz statt.
<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>	<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>

unterschiedliche Infos

Text 1:	Text 2:
Der Flieder blüht schon jetzt in wärmeren Regionen von Österreich. Das ist früher, als normal. Er blüht sogar schon mehr als eine Woche früher.	Der Flieder blüht in Österreich jetzt schon sehr früh. In den warmen Regionen ist der Flieder schon seit über einer Woche im Blühen.
<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>	<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>

links: Beistrichfehler, rechts: Grammatikfehler; unterschiedliche Aussagen

Text 1:	Text 2:
Die Staatsanwaltschaft machte eine Obduktion. Außerdem soll man herausfinden, ob die Buslenkerin etwas getrunken hat. Das heißt, es soll ein toxikologisches Gutachten geben.	Die Staatsanwaltschaft in Graz machte auch eine Obduktion und ein <b>toxisches</b> Gutachten. Das heißt, die Leiche von der Bus-Fahrerin wird untersucht.
<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>	<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>

rechts: richtig wäre „toxikologisch“

Text 1:	Text 2:
Es werde immer ein großes Bedauern für May sein, dass sie den Brexit nicht abgeschlossen bekommt. Das gestand sie ein.	Sie bedauerte, dass sie in der Lage gewesen sei, den Brexit durchzuführen.
<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>	<a href="#">DIESEN TEXT VERSTEHE ICH BESSER</a>

andere Aussage: links „nicht durchführen“, rechts „durchführen“; beides zu schwer

<p>Text 1:</p> <p>Die wurde am Donnerstag eingerichtet.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Diese Einrichtung wurde am Donnerstag eingerichtet. Dort arbeiten ein Tier-Arzt, ein Arzt und ein Anwalt.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	---

rechts: mehr Info, untypische Formulierung „Einrichtung eingerichtet“

<p>Text 1:</p> <p>Cannabis ist die am weitesten verbreitete Droge.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Die Droge, die auf der ganzen Welt am häufigsten konsumiert wird, heißt Cannabis.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
--	---

beides schwer

<p>Text 1:</p> <p>Er dürfte an den Folgen von dem Absturz gestorben sein, das glaubt die Bergrettung.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Der 23-Jährige aus dem Pinzgau ist an dem Unfall gewaltig gestorben. Das sagt die Bergrettung.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
---	--

links: Konjunktiv, nur 1 Satz; rechts: untypische Formulierung „gewaltig gestorben“

<p>Text 1:</p> <p>Die meisten Menschen ohne Internet-Anschluss leben in ärmeren Ländern.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>In ärmeren Ländern leben viele Menschen ohne einen Internetanschluss.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
--	---

nicht die gleiche Aussage

<p>Text 1:</p> <p>Der Chef von der Regierung rief die Oppositions-Parteien zu einem Misstrauens-Votum auf.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>	<p>Text 2:</p> <p>Der Regierungschef will die Oppositionsparteien zum Misstrauens-Votum herausfordern.</p> <p><b>DIESEN TEXT VERSTEHE ICH BESSER</b></p>
--	--

beides schwer

Text 1:  
  
Die Ausbildungsoffensive nutzt niemandem, wenn keine Jobs geschaffen werden.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
  
Die Ausbildungs- und Weiterbildungs-Offensive der türkis-grünen Regierung nutze nichts, wenn niemand einen Job bekommt.  
DIESEN TEXT VERSTEHE ICH BESSER

links ist deswegen leichter, weil Infos fehlen

Text 1:  
  
In Wullowitz in Oberösterreich sind am Montag zwei Messer-Attacken erfolgt.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
  
Am Montag hat es in der Gemeinde Wullowitz in Oberösterreich zwei Messer-Attacken gegeben.  
DIESEN TEXT VERSTEHE ICH BESSER

links: untypische Formulierung „Attacken erfolgt“

Text 1:  
  
Fleisch im Supermarkt ist in Österreich am teuersten in der EU.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
  
Österreich hat im EU-Vergleich am teuersten Fleisch.  
DIESEN TEXT VERSTEHE ICH BESSER

rechts: kein deutscher Satz

Text 1:  
  
Die EU-Kommissions-Chefin von der Leyen rechnet mit positivem Bescheid.  
DIESEN TEXT VERSTEHE ICH BESSER

Text 2:  
  
Von der Leyen rechnet mit einer positiven Entscheidung.  
DIESEN TEXT VERSTEHE ICH BESSER

links: „einem“ fehlt



## Appendix F

# Preference for Simple vs. Complex

Table F.1: **Preference for Simplified Text** (as opposed to complex text) among target and expert group participants. NA marks those target group annotators for which complex-vs-simple sanity check pairs were not annotated and expert group annotators who could see each pair's complex sentence in the web application. Only the ea03 value is statistically significant in a two-sided binomial test with a null hypothesis ratio of 0.5.

Target					
id	$\kappa$	id	$\kappa$	id	$\kappa$
ta01	0.500	ta06	NA	ta11	0.400
ta02	0.533	ta07	0.500	ta12	0.300
ta03	0.533	ta08	NA	ta13	NA
ta04	0.300	ta09	NA	ta14	0.692
ta05	0.400	ta10	0.400	ta15	NA

Expert			
id	$\kappa$	id	$\kappa$
ea01	0.477	ea03	<b>0.788</b>
ea02	NA	ea04	NA