Steering Large Language Models for Machine Translation Personalization

Anonymous ACL submission

Abstract

Large language models have simplified the production of personalized translations reflecting predefined stylistic constraints. However, these systems still struggle when stylistic requirements are implicitly represented by a set of examples, such as texts produced by a specific human translator. In this work, we explore various strategies for personalizing automatically generated translations when few examples are available, with a focus on the challenging domain of literary translation. We begin by determining the feasibility of the task and how style information is encoded within model representations. Then, we evaluate various prompting strategies and inference-time interventions for steering model generations towards a personalized style, with a particular focus on contrastive steering with sparse autoencoder (SAE) latents to identify salient personalization properties. We demonstrate that contrastive SAE steering yields robust style conditioning and translation quality, resulting in higher inference-time computational efficiency than prompting approaches. We further examine the impact of steering on model activations, finding that layers encoding personalization properties are impacted similarly by prompting and SAE steering, suggesting a similar mechanism at play.

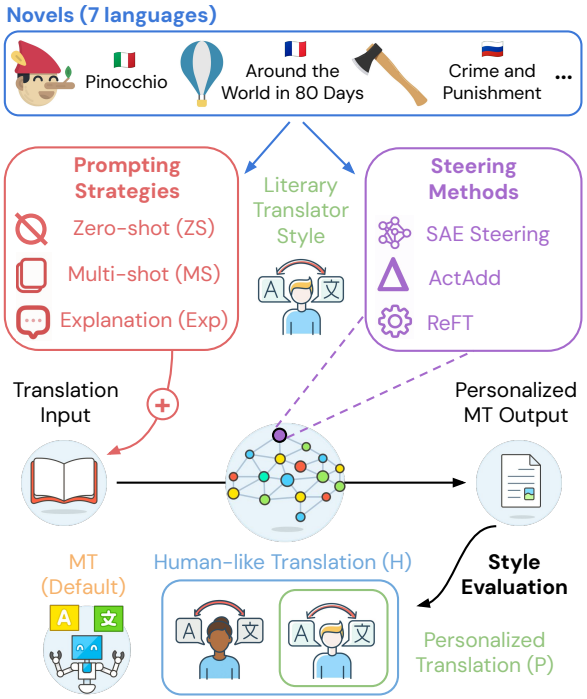

Figure 1: We compare prompt-based approaches with steering techniques intervening on model internals for personalizing MT outputs in literary machine translation. We use MT quality metrics and style classifiers to quantify the impact of steering on output fluency and personalization accuracy.

1 Introduction

A translated novel is not a simple literal retelling of the story—the personal style of the translator and their lexical and stylistic choices play a crucial role in shaping the narrative in a new language. Past efforts in the automatic translation of literary works have historically been constrained by the limited capabilities and flexibility of machine translation (MT) systems. The recent popularization of MT systems based on large language models (LLMs) has dramatically improved their capacity in handling the long contexts typical of literary translations, but mimicking the creative and rich language that characterizes the translators' own style remains an open issue. While prompting and tuning-based strategies have been successfully used for MT personalization (Michel and Neubig, 2018; Wang et al., 2021), recent advances in interpretability research have highlighted the effectiveness of various *steering* methods, which employ surgical interventions on LLMs' activations to condition model generation (Rimsky et al., 2024). While steering techniques were primarily applied to explicit properties such as sentiment and formality, this work aims to test their effectiveness for MT personalization. We compare steering methods with established prompting techniques, with a particular focus on the challenging domain of literary translation, where stylistic choices are more evident.

Provided that stylistic attribution is notoriously challenging for human annotators (Youyou et al., 2015; Flekova et al., 2016), we begin with a preliminary assessment to verify whether trained classifiers can distinguish between translators' styles and whether LLMs can use demonstrations to enhance the stylistic accuracy of their translations. We then connect the conditioning induced by prompting to the inner workings of the model, identifying activations with high discriminative capacity for style differences in intermediate layers of LLMs. In light of our results indicating that style information is effectively detected and represented by automatic systems, we evaluate the effectiveness of various established prompting and steering methods on three multilingual LLMs across novels in seven mid- to high-resourced languages. We focus in particular on a contrastive steering approach using sparse autoencoders (SAEs, Cunningham et al., 2023) to condition model generations by upweighting sparse style-related latents at inference time.

Our results demonstrate that contrastive SAE steering is a promising approach for MT personalization, yielding translations that align more closely with general human translation features and exhibit a greater similarity with the desired personalized style compared to other methods. Importantly, these results are achieved with no degradation in translation quality, assessed according to established MT quality metrics. We conclude by comparing the impact of our method on model representations with the outcome of multi-shot prompting, finding that probes trained on prompt-conditioned activations can predict the effectiveness of SAE steering with high precision. These results confirm that prompting and SAE steering techniques converge to similar solutions for conditioning model behavior, enabling future investigations into the mechanistic impact of prompting through the study of interpretable SAE latents.

2 Related Work

Machine Translation of Literary Texts The literary domain has historically been challenging for automatic MT systems due to their limited ability in handling rich linguistic and cultural contexts (Matusov, 2019) and their propensity to produce overly literal outputs (Guerberof-Arenas and Toral, 2022). Automatic literary translation has a long history dating back to pre-neural MT approaches (Voigt and Jurafsky, 2012; Toral and Way, 2015, 2018; Moorkens et al., 2018) with two recent dedicated evaluation campaigns (Wang et al., 2023b, 2024a). The advent of LLMs brought new opportunities in the processing of longer context for document-level translation (Wang et al., 2023a; Briakou et al., 2024; Wu et al., 2025a), but critical errors requiring human translator's intervention nonetheless persist (Karpinska and Iyyer, 2023). In this work, we use the PAR3 dataset (Thai et al., 2022), which comprises multiple human translations of novels to evaluate MT personalization in the literary domain.

Personalization for Machine Translation Advances in MT quality recently led to a growing interest in personalization approaches to ensure a consistent format and appropriate stylistic choices in model generations (Rabinovich et al., 2017; Lin et al., 2021). Previous techniques for controlling attributes such as formality (Sennrich et al., 2016; Niu et al., 2017; Nadejde et al., 2022) or gender (Vanmassenhove et al., 2018; Saunders and Byrne, 2020) typically required tuning existing models on pre-defined properties of interest, with few works attempting a real data-driven adaptation from unlabeled demonstrations (Michel and Neubig, 2018; Wang et al., 2021; Zhang et al., 2022). More recently, several studies employed prompting (Garcia and Firat, 2022; Sarti et al., 2023) or preference optimization from post-editing behavior (Lee et al., 2023; Berger et al., 2024) to render MT personalization more effective and data-efficient. In this work, we complement prompt results with steering approaches to personalize MT outputs using a few user-provided examples.

Steering Language Model Generations Steering approaches exploit the linear structure of LM activations (Mikolov et al., 2013; Chanin et al., 2024) to craft inference-time interventions for influencing model generations. These methods commonly employ contrastive sets of in-context demonstrations (Rimsky et al., 2024; Scalena et al., 2024) to map input properties to components such as vectors (Turner et al., 2024; Li et al., 2023), linear probes (Zou et al., 2025), or learned projections (Wu et al., 2024, 2025b). Sparse Autoencoders (SAEs) are another family of promising approaches for enabling fine-grained interventions in language models (Yun et al., 2021; Cunningham et al., 2023; Templeton et al., 2024). They are trained to decompose activations into approximately monosemantic features, offering a poten-

Lang.	Novel	Train / Val / Test
IT	Pinocchio	745 / 82 / 107
FR	Around the World in Eighty Days	829 / 92 / 120
NL	The Diary of a Young Girl	769 / 85 / 110
DE	Beware of Pity	606 / 67 / 96
RU	Crime and Punishment	1517 / 168 / 224
JA	No Longer Human	652 / 40 / 81
ZH	Dream of the Red Chamber	694 / 76 / 92

Table 1: Novels and number of example paragraphs used in our evaluation. Each example corresponds to a triple (MT, H1, H2). 20 additional examples per language are reserved for prompting/steering purposes.

ZH → EN 道人道：「既如此，便你去。」	
H1	*"Such being the case," the Taoist acquiesced, "I am ready to follow you, whenever you please to go."*
H2	*"Very good, I will go with you then," said the Taoist.*
Gemma 2 2B	*"If that's the case, then go ahead" said the Taoist.*
Gemma 2 9B	*"If so, then I will go with you" the Taoist said*
Llama 3.1 8B	*The Taoist said:"If you insist on going, go ahead."*
IT → EN Sarà l'effetto dell'acqua del mare. Il mare ne fa di questi scherzi	
H1	*"The salt water must have done it. The sea plays funny tricks."*
H2	*"It must have been the effect of sea-water. The sea makes extraordinary changes."*
Gemma 2 2B	*"It will be the effect of the sea water. The sea makes of these jokes."*
Gemma 2 9B	*It will be the effect of the sea water. The sea plays these tricks.*
Llama 3.1 8B	*It will be the effect of the sea water. The sea does things like this.*

Table 2: Examples from PAR3 for ZH → EN (*"Dream of the Red Chamber"* by Cao Xueqin) and IT → EN (*"The Adventures of Pinocchio"* by Carlo Collodi), including two human translations (H1, H2) and LLM outputs with zero-shot prompting. More examples in Appendix C.

tially interpretable basis for modifying model behavior. While interpreting their learned latents remains non-trivial (Marks et al., 2025), SAEs have proven effective for applying targeted interventions along specific linear directions (Chalnev et al., 2024; Zhao et al., 2025; Ferrando et al., 2025). However, most research on SAEs has so far focused on synthetic tasks or standard benchmarks, leaving their potential in real-world settings relatively underexplored.

3 Preliminaries

We begin our investigation by validating some key assumptions: **i)** Whether the personal translation styles are *discernible*, i.e., if the translation style of various LLMs and human translators can be consistently identified; **ii)** Whether LLMs can mimic specific translators' styles when provided with some of their translations; and **iii)** Whether style distinctions are reflected in the model's internal representations, to motivate the interest in steering approaches for improving MT personalization.

In our experiments, we use the PAR3 dataset by Thai et al. (2022), which comprises multiple human translations of novels from seven diverse languages (German, Russian, Chinese, Italian, Dutch, French, and Japanese) into English, as a benchmark for evaluating MT personalization. Novels are segmented into paragraphs with translations into English by two professional literary translators. Table 1 shows statistics about the PAR3 data, while Table 2 presents some examples for Chinese-to-English and Italian-to-English. We name the two available human translations H1 and H2, and compare them with MT outputs produced by LLMs, which we denote as MT_{model} from here onwards. We evaluate three LLMs, namely Llama 3.1 8B Instruct (Team, 2024b) and Gemma 2 (Team, 2024a) in its 2B and 9B instruction-tuned variants. Our model selection is motivated by our steering requirements, discussed in Section 5.

3.1 Is MT Personalization Discernible?

Following prior work on personalization (Wang et al., 2024c; Liu et al., 2023), we train a series of 📊 classifiers based on multilingual XLM Transformer encoders (Conneau et al., 2020) to distinguish between H1, H2, and MT translations for each language and each MT model in our evaluation suite. If classifiers can accurately predict the translation style among these three styles, then this provides evidence for distinct stylistic signals between them. In particular, the ability to distinguish between H1 and H2 would suggest that style signals can help differentiate between different human translators, rather than simply between human-like and automatic translations.[1] We find that the style information in the provided examples can be readily identified, with all 📊 classifiers reaching an accuracy between 77% for Japanese and 99% for Chinese, or 86% on average. These results corroborate previous evidence on the ability of automatic systems in detecting stylistic clues, while the task remains elusive for human annotators (Youyou et al., 2015; Flekova et al., 2016; Wang et al., 2024b).

[1]Details in Appendix A.2. Full results in Appendix B.1.

	Gemma 2 2B		Gemma 2 9B		Llama 3.1 8B	
ZS	0.10	0.69	0.08	0.71	0.08	0.70
MS	**0.24**	0.69	**0.31**	0.73	**0.32**	0.73

Table 3: Classifier-based personalization accuracy ( ) and Comet-based translation quality ( ) for zero-shot (ZS) and multi-shot (MS) prompting with 20 in-context examples averaged across all translators and languages.

3.2 Can LLMs Mimic Translation Styles?

We test the ability of LLMs to mimic the stylistic choices of a particular translator in a traditional multi-shot (MS) prompting setup. For each translator available across all tested novels, we prompt the model with 20 in-context examples selected from the original pool of translated paragraphs by that translator, asking it to generate new translations matching the provided style. We compare MS results with the default zero-shot (ZS) prompting using our high-scoring classifiers to quantify the effect of in-context examples on personalization accuracy, and using the popular Comet MT metric (Rei et al., 2020) to track translation quality. Table 3 presents our results. The proportion of outputs categorized as matching the translator's style is increased two- to four-fold when in-context examples are provided, confirming that LLMs can employ implicit clues in small sets of user examples to improve the style accuracy of their translations. Stable scores for Comet also indicate that translation quality is preserved after style adaptation.

3.3 Is Style Detectable in Model Activations?

In light of these results, we set out to test where the stylistic information enabling the improved performance of MS prompting is encoded within model representations. To this purpose, we train *linear probes* (Belinkov, 2022) using the last prompt token activations as input features to predict a binary class (MT or Human) and comparing it to the label the classifier (from Section 3.1) would assign to the resulting LLM translation.[2] High probing accuracy in this setting would suggest that the style information encoded from prompt examples is sufficient to predict the appropriateness of model translations before they are even generated. The training set of our probes comprises examples with 20 in-context demonstrations, where half of

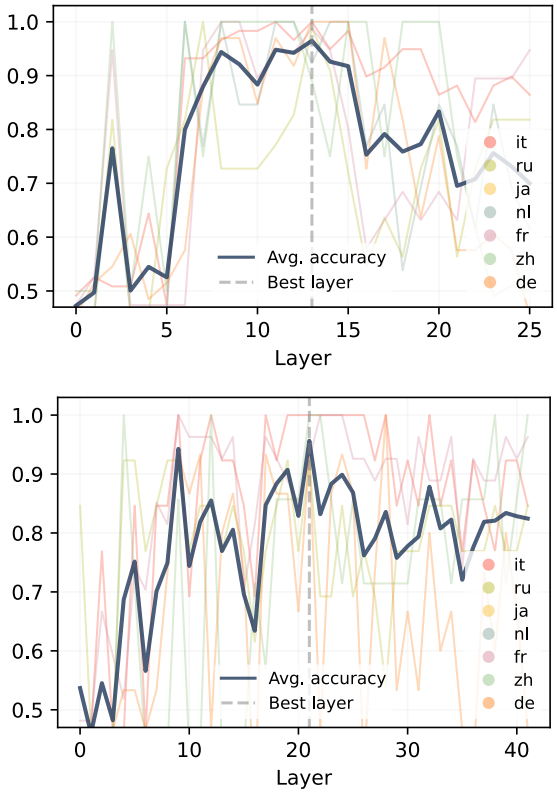

Figure 2: Probing classifier performance on the human translation detection task across Gemma 2 2B (top) and 9B (bottom) layers. Activations in intermediate layers are found to capture translation style information with high precision.

the examples use personalized translations made by a human translator, and the other half uses MT-generated demonstrations we pre-generated using the same model in a ZS setup. For each novel, we select as test examples only left-out entries for which the classifier predictions flip from MT in ZS setting to the desired style (H1or H2) when demonstrations are provided, to focus our analysis on settings where in-context demonstrations produce a tangible effect on output personalization.[3] This balanced setup precludes shortcut learning from spurious information to trained probes, e.g., the number of prompt demonstrations, ensuring that stylistic differences between human-made and MT-generated demonstrations are the sole distinguishing factor leading to discrepancies in model activations.

We focus specifically on Gemma models, extracting activations after the attention block at each model layer for the last token of the prompt, which was previously shown to encode key task-relevant

[2]To simplify the probing setup, the probe Human class is used for both H1 and H2 classifier predictions.

[3]In-context demonstrations are resampled for every test paragraph to prevent overfitting on spurious features.

information (Hendel et al., 2023; Todd et al., 2024; Scalena et al., 2024). Figure 2 reports probe accuracies across all Gemma 2 2B layers, with results for the 9B model reported in Appendix B. We find a peak in probe accuracy of ∼95% around intermediate model layers, suggesting that these layers encode stylistic information with very high precision.[4] These results confirm that personalization features are clearly discernible from LLMs' activations, motivating our following experiments on activation steering for MT personalization.

4 Methods

We now introduce the methods evaluated in our experiments, with a focus on our proposed SAE-based contrastive steering method. In all our contrastive formulations, we treat the available H1 and H2 translations as distinct personalization targets to contrast with MT, and use H^+, H^- to indicate desired and undesired human styles, respectively.[5]

4.1 Prompting Baselines

Zero-Shot (ZS). The ZS setup of our main experiment matches the one from Section 3.1, in which the model is asked to produce a translation with no conditioning from examples or explanations towards the target translation style. We use this setting to establish a baseline style and translation quality performance for tested models.

Zero-Shot with Explanation (ZS-Exp). We extend the ZS setting by exploiting verbalized style description produced by GPT-4o (OpenAI, 2024). We prompt GPT-4o with 20 H^+ demonstrations and 20 MT outputs produced in the ZS settings, asking it to synthesize a detailed explanation of salient elements characterizing the desired style.[6] We manually review all generated explanations to ensure they do not contain any verbatim excerpts from input examples, and use them to the tested LLM in a setting with no demonstrations (ZS). This setup allows us to verify whether natural language directives can act as effective and interpretable alternatives to other tested methods.

Multi-Shot (MS). Following Section 3.2's findings, we adopt the same MS setup using 20 in-context translation examples matching the style of a desired human translator H^+.

[4] We find probes for layers 13 and 21 to perform best for the 2B and 9B models, respectively.

[5] E.g., when mimicking H1 style, H^+= H1 and H^-= H2.

[6] Prompt templates in Listings 1, 2,3 of Appendix A.5.

4.2 Steering Baselines

Activation Addition (ActAdd). ActAdd is a simple yet effective technique for steering language models. We employ the standard formulation by Rimsky et al. (2024), contrasting a set of *positive* activations $\{z\}^+$, here the style-aligned H^+ translations, with *negative* activations $\{z\}^-$, here the default MT output. We perform steering on the style-relevant layers identified in Section 3.3 by computing the average Δ steering vector as the difference-in-means for the two sets of activations using 20 in-context examples, and applying it additively to the same model layer during inference. We use a factor $\alpha = 2$ to control steering intensity, as this was found effective in previous research (Scalena et al., 2024).

Representation Fine-tuning (ReFT). ReFT was recently introduced as an alternative to traditional weight-based Parameter-Efficient Fine-Tuning (PEFT) techniques, aiming to learn task-specific interventions applied directly to model activations at inference time (Wu et al., 2024). As for ActAdd, we apply ReFT to the style-relevant layers identified in Section 3.3 and limit confounding factors by tuning ReFT interventions with the same set of 20 examples used for MS prompting.

4.3 Contrastive SAE Steering (SAE Cont.)

We now introduce our proposed contrastive formulation for steering translation style with SAE latents. Given the set of paragraphs $\mathcal{D}$ for a novel in the PAR3 dataset, each instance in it is a tuple $\langle s, H^+, H^-, \text{MT}\rangle$, where s is the non-English source sentence. Similar to previous methods, we employ a contrastive approach to extract SAE latents that are most active in the presence of the selected translator's style, while simultaneously controlling for generic features occurring in natural model outputs.

Feature extraction First, we gather activations z_l^+ and z_l^- by prompting the model with tuples $\langle s, H^+\rangle$, $\langle s, \text{MT}\rangle$, respectively. Activations are extracted at the last prompt token position from the most style-related layer, as identified in Section 3.3. Activations are then converted into sparse feature dictionaries $x^+ = h(z^+)$ and $x^- = h(z^-)$, with $x^+, x^- \in \mathbb{R}^m$ and h representing the the SAE encoder module. This procedure is repeated separately across 20 contrastive examples, resulting in two collections of SAE latent vectors for posi-

tive/negative examples, $\mathcal{X}^+ = \{x_1^+, x_2^+, \ldots, x_{20}^+\}$ and $\mathcal{X}^- = \{x_1^-, x_2^-, \ldots, x_{20}^-\}$

Relevance-based Feature Selection To identify discriminative features for personalization in the large set of latents, we adapt the information-theoretic approach proposed by Zhao et al. (2025). For each of the inputs, we identify the subset of size $n \ll m$ of SAE active features, i.e. latent dimensions for which the logit is > 0. We consider logit values in this subset as instances of a random variable $X_i \in x$, and calculate the mutual information $I(X_i, Y)$ between each feature X_i and the target binary variable $Y = \{+, -\}$, representing whether the text matches the desired H^+ style or not. A higher $I(X_i, Y)$ indicates that the i-th feature is more informative for discriminating between personalized and default inputs, and can hence be used for steering. A representative sample of 40 latents showing the highest mutual information scores for both personalized ($\{X_i\}^+$) and non-personalized ($\{X_i\}^-$) examples is selected using this procedure. This procedure differs from traditional SAE-based steering methods, which employ only features associated with the positive class (Chalnev et al., 2024; Arditi et al., 2024), in that it encourages the selection of discriminative features, both positive and negative. For every selected latent, we use its expected logit when personalization is present or absent, given the set of provided examples, i.e. $\mathbb{E}^+[X_i]$ and $\mathbb{E}^-[X_i]$.

Inference-time intervention Finally, activations are steered by clipping selected latents' values to their expected value, hence promoting H^+-aligned latents $\{X_i\}^{H^+}$ while demoting MT-aligned ones $\{X_i\}^{\text{MT}}$.[7] Additionally, we modulate the magnitude of the resulting vector with an α *coefficient*, which was found to play an essential role in steering effectiveness in previous research (Scalena et al., 2024; Ferrando et al., 2025)

5 Experiments

5.1 Setup

Model selection We evaluate our methods on the Llama 3.1 and Gemma models previously used in Section 3. Our selection is guided by the availability of open-source pre-trained SAEs, which can be otherwise computationally expensive to train. For Gemma models, we employ SAEs from the GemmaScope suite (Lieberum et al., 2024), while for the Llama 3.1 model we use the SAE by McGrath et al. (2024). GemmaScope SAEs are available for every model layer, enabling us to steer Gemma models on their most informative layers for the task, which we identified in Section 3.3. On the contrary, a single SAE for the 19th layer is available for Llama, hence limiting our evaluation of SAE steering and potentially producing sub-optimal steering results for that model.

Metrics We evaluate our approaches on a held-out test set sourced from the PAR3 dataset. For assessing style accuracy, we use the classifiers described in Section 3.1. We define three submetrics employing the classifier probability distribution over the three classes (MT, H1, H2) to better analyze different aspects of model outputs. First, we define the **H** accuracy as the classifier's total probability assigned to human-like translations, $p(\text{H1}) + p(\text{H2})$, thereby measuring generic *human-like* features of the text. We instead define **P** accuracy as the probability of the desired H^+ class, which corresponds to specific stylistic traits of a human translator. Finally, we use $\mathbf{P}_{\text{FLIP}}$ to indicate the proportion of examples for which the classifier prediction flips from MT to H^+ after prompting or steering is applied, pinpointing examples for which the method has an observable effect on generation style. To ensure that our interventions do not result in a degradation of overall translation quality, we also employ Comet[8] (Rei et al., 2020) using the H^+ as reference.

The α trade-off We begin by verifying the optimal steering intensity α for our SAE steering technique. We primarily focus on results from Gemma 2 2B, for which we ran a comprehensive sweep over all relevant hyperparameters.[9] Figure 3 (top) shows the influence of α on MT personalization accuracy and fluency, averaged across all translators and all tested languages. For values of $\alpha \leq 3$, performance remains close to that of the MS baseline, indicating that the contrastive method is effectively isolating latents associated with the desired style. As α increases, performance generally exceeds the MS approach, achieving greater stylistic accuracy with minimal impact on translation quality. However, for $\alpha \geq 10$, we observe a degradation in translation fluency, represented by Comet.

[7] Algorithm 1 summarizes our SAE steering approach.

[8] Unbabel/wmt22-comet-da

[9] Larger models were evaluated using a subset of the best-performing configurations. Details in Appendix B.

	Gemma 2 2B					Gemma 2 9B					LLaMA 3.1 8B				
	H	**P**	**P** $_{\text{FLIP}}$		tok/s	**H**	**P**	**P** $_{\text{FLIP}}$		tok/s	**H**	**P**	**P** $_{\text{FLIP}}$		tok/s
ZS	0.21	0.10	0.05	0.69	39.8	0.15	0.08	0.04	0.71	**25.7**	0.24	0.08	0.05	0.70	**25.9**
ZS-Exp.	0.30	0.22	0.16	0.68	**41.3**	0.41	0.22	0.18	0.72	24.6	0.56	0.23	0.21	0.69	25.5
MS	0.37	0.24	0.16	0.69	36.0	**0.48**	0.31	0.27	**0.73**	16.8	0.58	0.32	**0.28**	**0.73**	17.2
ActAdd	0.27	0.22	0.12	0.67	40.2	0.32	0.24	0.20	0.70	25.3	0.55	0.36	**0.28**	0.70	24.2
ReFT	0.31	0.22	0.18	**0.70**	40.7	0.46	0.34	0.27	0.67	25.5	0.53	**0.38**	0.26	0.70	24.7
SAE Cont.	**0.39**	**0.27**	**0.19**	**0.70**	41.1	0.46	0.33	**0.29**	0.72	25.1	**0.59**	0.31	0.27	0.72	23.1

Table 4: Averaged metric scores across all tested languages (per-language breakdown in Appendix B). **H**: human style accuracy, i.e. $p(\text{H1}) + p(\text{H2})$. **P**: personalization accuracy $p(H^+)$ for the desired style. **P** $_{\text{FLIP}}$: % of segments for which style conditioning flips the classifier prediction from MT to H^+. $\alpha = 5$ is used for SAE Cont. results.

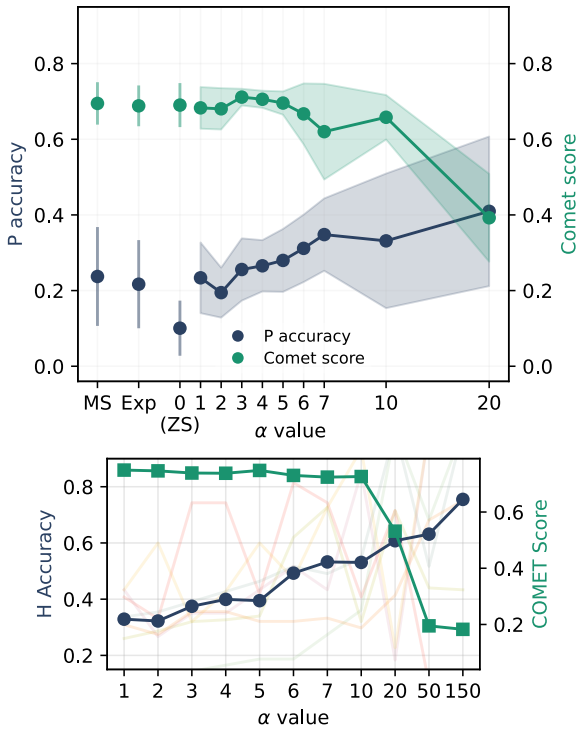

Figure 3: Effect of various steering intensity α on style accuracy and translation quality for Gemma 2 2B. **Top:** **P** accuracy for SAE Cont. and prompting baselines (MS, ZS-Exp and ZS). **Bottom:** H accuracy for high α showing a steep drop in translation quality while style accuracy increases.

Following the SAE steering setup of Ferrando et al. (2025), we also experiment with very high α (up to 150), obtaining **H** accuracies approaching 100% for some languages but also resulting in abysmal translation quality, with generally nonsensical outputs (Figure 3 bottom). These results suggest that SAE features identified by our contrastive approach correspond to human-like constructions, provided that their upweighting raises the classifier confidence in predicting H1 or H2, even when the resulting quality is clearly lacking.[10] We leave an exploration on using such examples to further improve style classifiers to future work. Ultimately, we identify $\alpha = 5$ as an appropriate steering intensity that balances personalization and fluency, and employ it in our main evaluation.

[10]A qualitative evaluation is provided in Appendix C.

5.2 Results and Discussion

Table 4 presents the performance of tested models across prompting and steering setups, averaged across all languages and styles, i.e., using H1 and H2 as H^+ for each language. We find that contrastive SAE steering (SAE Cont.) yields a good balance between style accuracy and translation quality, with results comparable to the strong MS prompting baseline. Notably, SAE Cont. outperforms all other tested methods for the smaller Gemma 2 2B model. We conjecture this could be due to the larger models' ability to incorporate in-context information more effectively, reducing the benefits of ad-hoc interventions.

Is steering computationally efficient? To quantify the impact of various steering strategies on inference speed, for each method in Table 4 we include the resulting generation speed in tokens per second (tok/s) estimated over the whole test set. Despite their comparable performances, we observe that MS prompting is systematically slower than SAE Cont. steering across all tested models when using 20 in-context demonstrations. We set out to test the trade-off between style conditioning and computational efficiency, focusing on the translation direction with more paragraphs available (Russian→English) as we vary the number of examples used in prompting and steering. Our results in Figure 4 show that increasing the number of examples used to identify features yields consistent improvements for SAE Cont., while scaling in MS prompting is limited by the model's limited context size. Moreover, the inference-time cost of SAE Cont. steering remains constant when increasing the demonstrations used for selecting steered fea-

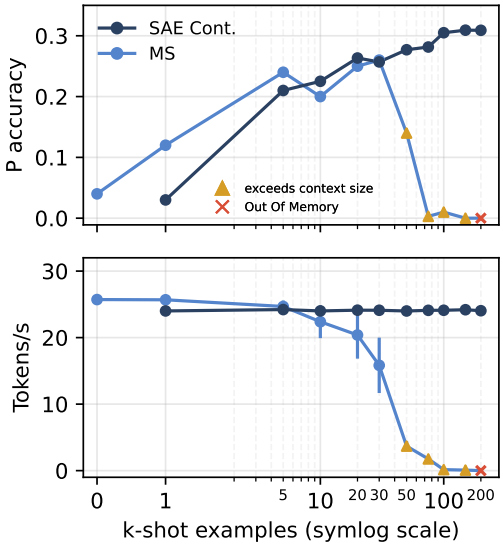

Figure 4: Personalization accuracy **P** (top) and inference speed (Tokens/s) (bottom) across in-context demonstration counts, using Gemma-2 9B for Russian → English translation. More results in Appendix E.

tures, while MS prompting undergoes a significant slowdown when demonstrations are added to the prompt. Overall, these results strongly highlight how the SAE Cont. method, while comparable to MS prompting in smaller setups, can result in more scalable performances without requiring additional inference time computations. We report additional evidence in Appendix E.

Is SAE steering an "internalized" MS prompting? In light of the comparable results obtained by SAE steering and MS prompting, we set out to investigate whether the two methods also modify model activations in a similar way. For this purpose, we use the *probing classifiers* introduced in Section 3.3, which were trained on MS-conditioned last prompt token activations for predicting the style of the resulting text. We categorize test examples in three groups: one in which steering flips the original classifier prediction to a human label, i.e. H1 or H2 (MT → H), and two in which prediction remains unchanged (MT →MT, $H \rightarrow H$). For each example, we collect steered activations z_{steer} using the SAE Cont. method and calculate the average probability of human-like output (H) predicted by the probe over each category. Our results in Table 5 show that, for both Gemma models, probes can predict successful (MT → H) and unsuccessful (MT →MT) steering outcomes from SAE-steered activations, despite being trained only on MS prompting examples. In settings where

	MT → H	MT → MT	$H \rightarrow H$
Gemma 2 2B	0.94 ±0.01	0.07 ±0.02	0.72 ±0.15
Gemma 2 9B	0.93 ±0.02	0.12 ±0.10	0.68 ±0.19

Table 5: Estimated $p(H)$ for SAE-steered activations using Section 3.3 binary probes trained on MS-conditioned activations, across three settings identified by classifier input/output labels.

the original output already matches human style ($H \rightarrow H$), instead, probes reach lower accuracy with broader confidence intervals, denoting higher uncertainty. The ability of probes to generalize to SAE steering from prompting examples provides strong evidence that **SAE Cont. steering mirrors the effect of in-context demonstrations from MS prompting on model activations**. This suggests, in turn, that SAE features identified by our contrastive setup might act as efficient "internalized summaries" of prompt demonstrations.

6 Conclusion and future work

We conducted a broad evaluation of various prompting and steering approaches for personalizing LLM-generated translations. Our evaluation targets a practical, real-world application of literary translation, addressing the underexplored challenge of steering LLM generations in a linguistically rich and stylistically sensitive domain. Through comprehensive evaluation across multiple languages, novels, and models, we demonstrate that our proposed SAE-based approach outperforms prompting and alternative steering techniques, resulting in a good balance of style conditioning and translation quality at negligible inference-time cost. In particular, our proposed method scales effectively to large amounts of examples, bypassing the context limitations and slowdowns of in-context demonstrations.

Although faithfully replicating individual human translation styles remains a highly challenging task, our approach achieves strong alignment with human translation quality and translator-specific style features. In particular, the effectiveness of our proposed approach on smaller models might enable MT personalization in smaller-scale computational settings, easing further research on how style information is encoded and produced by language models.

Limitations

While our work demonstrates the potential of steering LLMs for MT personalization using sparse au-

toencoders, we acknowledge some limitations in our setup.

Firstly, the generalizability of our findings is constrained by the scope of our experiments. We focused on literary translation from seven specific source languages into English and evaluated three LLMs of relatively small size. Consequently, the effectiveness of SAE-based steering and the identified optimal layers for intervention may not directly transfer to other language pairs, significantly different model architectures or sizes, or distinct domains beyond literary texts. Further research is needed to assess the robustness of our approach across a broader range of linguistic and modeling contexts.

Secondly, the computational overhead associated with training sparse autoencoders presents a practical limitation to the widespread adoption of our proposed method. Although many open-source pre-trained SAEs are now available, training new SAEs from scratch may be required for newer models, which might pose a problem for researchers with limited computational resources. The current availability of pre-trained SAEs also restricts model choice, as seen with the Llama 3.1 8B model where the SAE we selected was available for a single, potentially suboptimal layer.

Finally, our investigation primarily focused on downstream performance and the impact of various personalization strategies on model activations. However, we did not pursue a mechanistic understanding of the "personalization circuits" within the LLMs. Future work could adopt a more fine-grained, mechanistic interpretability approach to study how specific SAE latents or combinations thereof encode and manipulate nuanced stylistic features, thereby providing deeper insights into the underlying processes of LLM personalization.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Post-edits are preferences too. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1289–1300, Miami, Florida, USA. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features.

David Chanin, Anthony Hunter, and Oana-Maria Camburu. 2024. Identifying linear relational concepts in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1524–1535, Mexico City, Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models.

Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, and David Bau. 2024. Nnsight and ndif: Democratizing access to foundation model internals.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation. *Translation Spaces*, 11(2):184–212.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Jihyeon Lee, Taehee Kim, Yunwon Tae, Cheonbok Park, and Jaegul Choo. 2023. PePe: Personalized post-editing model utilizing user-generated post-edits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 239–253, Dubrovnik, Croatia. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4008–4018, Online. Association for Computational Linguistics.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023. CoCo: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188, Singapore. Association for Computational Linguistics.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Thomas McGrath, Daniel Balsam, Myra Deng, and Eric Ho. 2024. Understanding and steering llama 3 with sparse autoencoders.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 577–603, Miami, Florida, US. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size.

Llama Team. 2024b. The llama 3 herd of models.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*. ArXiv:2310.15213.

Antonio Toral and Andy Way. 2015. Translating literary text between related languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 123–132, Denver, Colorado, USA. Association for Computational Linguistics.

Antonio Toral and Andy Way. 2018. *What Level of Quality Can Neural Machine Translation Attain on Literary Text?*, pages 263–287. Springer International Publishing, Cham.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024a. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Yue Wang, Cuong Hoang, and Marcello Federico. 2021. Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1193–1199, Online. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025a. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025b. Axbench: Steering llms? even simple baselines outperform sparse autoencoders.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. ReFT: Representation fine-tuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online. Association for Computational Linguistics.

Peng Zhang, Zhengqing Guan, Baoxi Liu, Xianghua (Sharon) Ding, Tun Lu, Hansu Gu, and Ning Gu. 2022. Building user-oriented personalized machine translator based on user-generated textual content. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2025. Representation engineering: A top-down approach to ai transparency.

A Experimental Reproducibility

In this section, we provide every parameter we use for the reproducibility of our experiments setups.

A.1 Base prompt

We use the same prompt template across all methods: ZS (which corresponds to the original model translation), ZS-Exp. (detailed in Appendix A.5), MS, ActAdd, ReFT, and SAE Cont. This prompt, shown in Listing 3, instructs the model to translate the source sentence while explicitly preventing it from adding any explanations about the translation process. Since all test models are Instruction Tuned, we utilize their native chat templates to preprocess the input accordingly. For multi-shot examples, the *user* and *assistant* turns are repeated for each example, always using the same prompt structure.

A.2 Classifier training

All classifiers are fine-tuned from the `xlm-roberta-large` model[11], using a linear classification head. Training is conducted for 6 epochs with a learning rate of 2e-5 and a batch size of 32, selecting the best model checkpoint based on validation accuracy.

Training data only includes generations from models and the translator without any source text. It is also perfectly balanced, as each paragraph provides one instance for all three labels: H1, H2, and MT. The total size of the training set varies depending on the number of paragraphs in the chosen novel. On average, we obtain approximately 830 instances, resulting in a total of around 2,490 labeled examples for training (see 1). Validation and test sets are strictly held out and never seen during training. Additionally, they do not include the small 20-example subsets used in the MS, ZS-Exp. and SAE Cont. setups.

A.3 ReFT training

ReFT training was conducted using the PyReFT toolkit from the original authors[12]. We applied the intervention at the same hook point used by other steering methods - specifically, the layer output corresponding to the residual stream at the selected layer. The training configuration includes a `low rank dimension` of 4, `lora alpha` set to 32, and a `lora dropout` of 0.05. ReFT was trained on the same 20 prompts used in the MS setup, for a total of 100 epochs.

A.4 SAE Cont.

We use the NNsight library (Fiotto-Kaufman et al., 2024) to extract and manipulate model activations for all steering experiments. The source code is publicly available in the repository linked in the main body of this paper. For consistency, we use the same set of contrastive examples employed in the MS approach.

Algorithm 1 outlines the procedure for latent-based steering. It enhances features identified as relevant to personalization while simultaneously suppressing those negatively correlated with the task.

Algorithm 1: Contrastive SAE Steering

Input: Input activation z, SAE model `sae`, target latents expected value $\mathbb{E}^+[X_i]$, contrast latents expected value $\mathbb{E}^-[X_i]$, steering coefficient α
Output: Steered activation z_{new}
$x \leftarrow \texttt{sae.encode}(z)$;
$m \leftarrow \text{length}(x)$;
for $i \leftarrow 1$ **to** m **do**
  if $\mathbb{E}^+[X_i] > x[i]$ **then**
    $x[i] \leftarrow \mathbb{E}^+[X_i]$
  if $\mathbb{E}^-[X_i] < x[i]$ **then**
    $x[i] \leftarrow \mathbb{E}^-[X_i]$
$z_{\text{steer}} \leftarrow \alpha \cdot \texttt{sae.decode}(x)$;
return z_{steer}

A.5 ZS-Exp.

For ZS-Exp., we used GPT-4o (June 2025) to generate explanations detailing the stylistic differences between a base translation and a target human translation. The prompt template used for this task is shown in Listing 1, using the same 20 examples as in the MS and SAE Cont. setups. Example outputs for different novels are shown in Listing 2. Finally, these generated guidelines are used to prompt the evaluated models, following the template shown in Listing 3.

B Results across languages

B.1 Classifiers

We show in Table 6 results for every 📊 classifier trained for each model and for each language.

B.2 Prompting and steering results

We present detailed plots of the results for each novel across the three evaluated models in Figure 6 (Gemma 2 2B), Figure 7 (Gemma 2 9B),

[11] FacebookAI/xlm-roberta-large
[12] `stanfordnlp/pyreft`

Listing 1: Prompt template used to get GPT 4o explanation using translation examples.

```
Objective: Identify stylistic choices in translations for personalization purposes.

You will be provided with a source text, a standard translation, and a target translation by a specific
    translator whose style we want to emulate.
Your task is to analyze the 'Target translation' by comparing it to the 'Base translation' and the 'Source
    text'.
Identify and list the distinctive stylistic patterns, choices, and preferences exhibited in the Target
    translation.

These stylistic cues should help another translator (or an AI) to adapt their translations to match the
    style of the target translator.

Source text: <source text here>
Base translation: <MT  text here>
Target translation: <H+ translation here>

<... repeat Source, MT and Target>

Please extract a concise list of key stylistic cues. Focus on aspects such as vocabulary choices, sentence
    structure, tone and register, handling of cultural nuances, punctuation/formatting preferences and
    overall creativity.

Output a short list of stylistic cues as bullet points. Write the list as if you were directly giving the
    guidelines to the translator and avoid using specific examples.
```

Lang.	Gemma 2 2B	Gemma 2 9B	Llama 3.1 8B
DE	0.89	0.90	0.84
RU	0.92	0.90	0.91
ZH	0.99	0.98	0.98
IT	0.78	0.85	0.80
NL	0.79	0.78	0.82
FR	0.88	0.87	0.90
JA	0.76	0.79	0.76

Table 6: Accuracy of model- and language-specific 3-way (MT, H1, H2) classifiers on balanced held-out sets for every language. Random baseline: 0.33.

and Figure 8 (Llama 3.1 8B). These plots display the performance of all evaluated methods, reporting the three submetrics: **H** accuracy (general human-likeness), **P** accuracy (translator-specific accuracy), and **P** $_{\text{FLIP}}$ (personalized flip accuracy), alongside the corresponding Comet scores measuring translation quality.

C Examples from Dataset and Conditioning Outputs

We present in Tables 8 and 9 a selection of examples from two different languages, showcasing outputs from each of the tested setups. For each example, we also report the corresponding classification label predicted by the classifier and the associated Comet score. Additionly table 10 shows some examples of models generating output aligned with the Human translator according to the classifier but with a low Comet score corresponding to an almost unreadable output due to extreme α values.

D Additional Experiment on Contrastive Setups to Isolate Style Features

To isolate features relevant to a specific translation style, we adopt a contrastive approach between different parallel translation examples. Specifically we test two setups:

- $M \leftrightarrow H$: contrasting the target style we aim to achieve (H^+) with a "clean" style derived from the model's original machine translation (M);
- $H_\alpha \leftrightarrow H_\beta$ contrasting the desired style with the style of another human translator.

The $M \leftrightarrow H$ setup is designed to extract features that distinguish the target human-like style from a standard, model-generated translation. The $H_\alpha \leftrightarrow H_\beta$ setup, on the other hand, aims to isolate the features that differentiate one translator's style from another. While the latter more closely

Listing 2: Examples of explanation obtained from GPT 4o when comparing different translations from different novels.

When comparing H1 and MT for Beware of Pity (German):
- Maintain a tone that is professional, thoughtful, and subtly persuasive.
- Avoid overly technical jargon unless necessary; explain specialized terms briefly if used.
- Preserve the author's voice, keeping a balance between academic rigor and narrative engagement.
- Ensure smooth transitions between sentences and paragraphs to support coherent argumentation.
- Translate idiomatic expressions in a way that retains their intended effect, even if the wording differs.

When comparing H1 and H2 for Beware of Pity (German):
- Keep the tone warm, welcoming, and direct – avoid overly formal or distant language.
- Use active voice wherever possible to maintain energy and engagement.
- Avoid idioms or expressions that may not translate culturally; aim for universal accessibility.
- Maintain consistent tone and register throughout, adapting to the intended audience's familiarity with the subject.
- Respect the rhythm and structure of the original, but feel free to adjust sentence length for readability.

When comparing H1 and H2 for Crime and Punishment (Russian):
- Prefer dynamic over formal vocabulary: Opt for vivid or emotionally charged words when available.
- Add tonal nuance and emotional shading: Enrich dialogues and narration with subtle shifts in tone, especially sarcasm, understatement, or irony, to match character voice or mood.
- Use contractions and familiar phrasing: Employ contractions and relaxed expressions to preserve spoken character.
- Expand or rephrase for clarity and voice: Don't hesitate to slightly reword or elaborate if it strengthens tone, clarifies intent, or enhances character differentiation.
- Favor rhythmic, flowing sentence structure: Break long, formal sentences into multiple shorter clauses or use punctuation (dashes, ellipses) for dramatic or emotional effect.
- Reflect subtle character dynamics: Infuse lines with interpersonal undertones (like defiance, deference, or sarcasm) that may not be explicit in the original.
- Preserve or recreate emotional tension: Use word choice and pacing to sustain psychological nuance, unease, or irony.
- Use expressive punctuation and formatting: Favor dashes, ellipses, and italic like emphasis (through word placement) to reflect emotional cadence or interruptions.

corresponds to the notion of translation style (as a property unique to each translator), we chose to experiment with both since prior works have shown that contrasting the desired behavior with the model's baseline behavior tends to yield the most effective results when steering LLMs.

Which contrastive setup is better? We present in Table 7 the results for both contrastive setups. For both the ZS-Exp. and SAE Cont. approaches, the results are generally comparable, making it difficult to identify a clear best-performing method. However, larger models occasionally achieve higher performance, which we hypothesize may stem from their enhanced ability to disentangle personalization-relevant features even without explicit guidance.

E Additional Results when Scaling Demonstrations Count

Figure 5 presents the full set of experiments, including two languages (Russian and French) and two models (Gemma 2 2B and 9B), complementing the results of Figure 4. The two languages were selected as they presented the largest collection of translated paragraphs available (see Table 2), allowing us to scale the ICL set without substantially reducing the size of the training data or impairing classifier performance. Notably, as also shown in Figure 4 the personalization accuracy **P** of our contrastive approach improves across all models and language tested, even surpassing the multi-shot (MS) baseline. In all cases, the MS baseline reaches the LLM context size limit (around ICL ≥ 50), after which the performance degrades

Listing 3: Zero shot template template when prompting language models with different setups

```
Translate the following sentence between the angular parentheses into English.
if setup == ZS−Exp
{
   Follow the following guidelines when translating:
   <explanations here>
}

The original sentence is: <source text>.

Remember to write only the translation, without any additional text or explanation.
```

Method	H	P	P_{FLIP}		tok/s
Gemma 2 2B					
ZS-Exp$_{M \leftrightarrow H}$	0.30	0.22	0.16	0.68	**41.3**
ZS-Exp$_{H_\alpha \leftrightarrow H_\beta}$	–	0.20	0.14	0.69	41.2
SAE Cont.$_{M \leftrightarrow H}$	**0.39**	**0.27**	**0.19**	**0.70**	41.1
SAE Cont.$_{H_\alpha \leftrightarrow H_\beta}$	–	**0.27**	0.18	0.69	40.8
Gemma 2 9B					
ZS-Exp$_{M \leftrightarrow H}$	0.41	0.22	0.18	0.72	24.6
ZS-Exp$_{H_\alpha \leftrightarrow H_\beta}$	–	0.23	0.19	**0.73**	**24.4**
SAE Cont.$_{M \leftrightarrow H}$	**0.46**	0.33	**0.29**	0.72	25.1
SAE Cont.$_{H_\alpha \leftrightarrow H_\beta}$	–	**0.35**	**0.29**	**0.73**	24.9
LLaMA 3.1 8B					
ZS-Exp$_{M \leftrightarrow H}$	0.56	0.23	0.21	0.69	25.5
ZS-Exp$_{H_\alpha \leftrightarrow H_\beta}$	–	0.30	0.26	0.70	25.5
SAE Cont.$_{M \leftrightarrow H}$	**0.59**	0.31	0.27	**0.72**	**23.1**
SAE Cont.$_{H_\alpha \leftrightarrow H_\beta}$	–	**0.33**	**0.28**	**0.72**	23.6

Table 7: Averaged metric scores across all tested languages for the $M \leftrightarrow H$ and $H_\alpha \leftrightarrow H_\beta$ variants of ZS-Exp. and SAE Cont. methods. **H**: human style accuracy, i.e. $p(\text{H1}) + p(\text{H2})$. **P**: personalization accuracy $p(H_x)$ for the target style. $\mathbf{P}_{\text{FLIP}}$: Proportion of segments for which steering has a causal impact on personalization. $\alpha = 5$ is used for SAE Cont. results.

significantly. Similarly, the tok/s efficiency metric highlights the advantage of our contrastive approach: It brings only a minimal, roughly linear, decrease in inference-time speed (compared to ZS, i.e. ICL $= 0$), whereas the MS setup exhibits a progressively larger computational impact as the number of in-context examples increases.

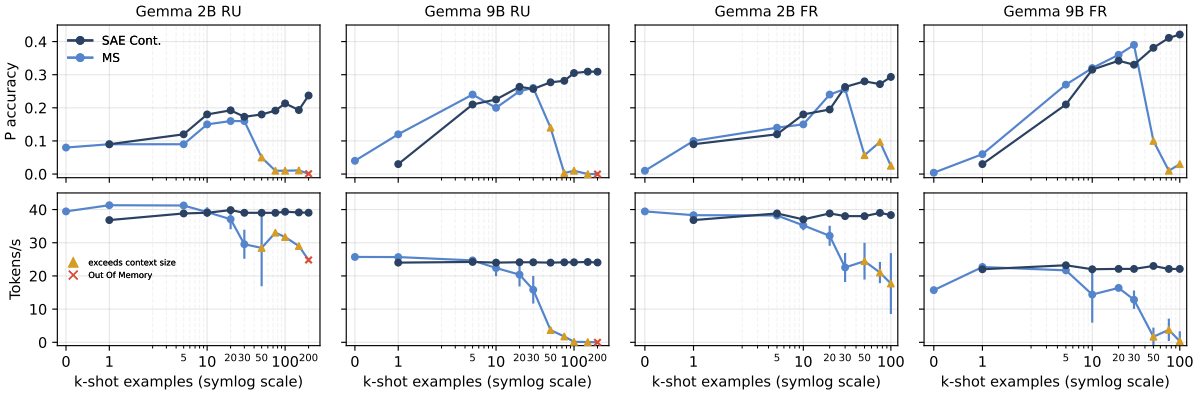

Figure 5: Complete results when comparing the MS approach to our SAE Cont.$_{M \leftrightarrow H}$ for the Gemma models (2B and 9B) on the largest novels, evaluated at the paragraph level, in Russian (RU) and French (FR).

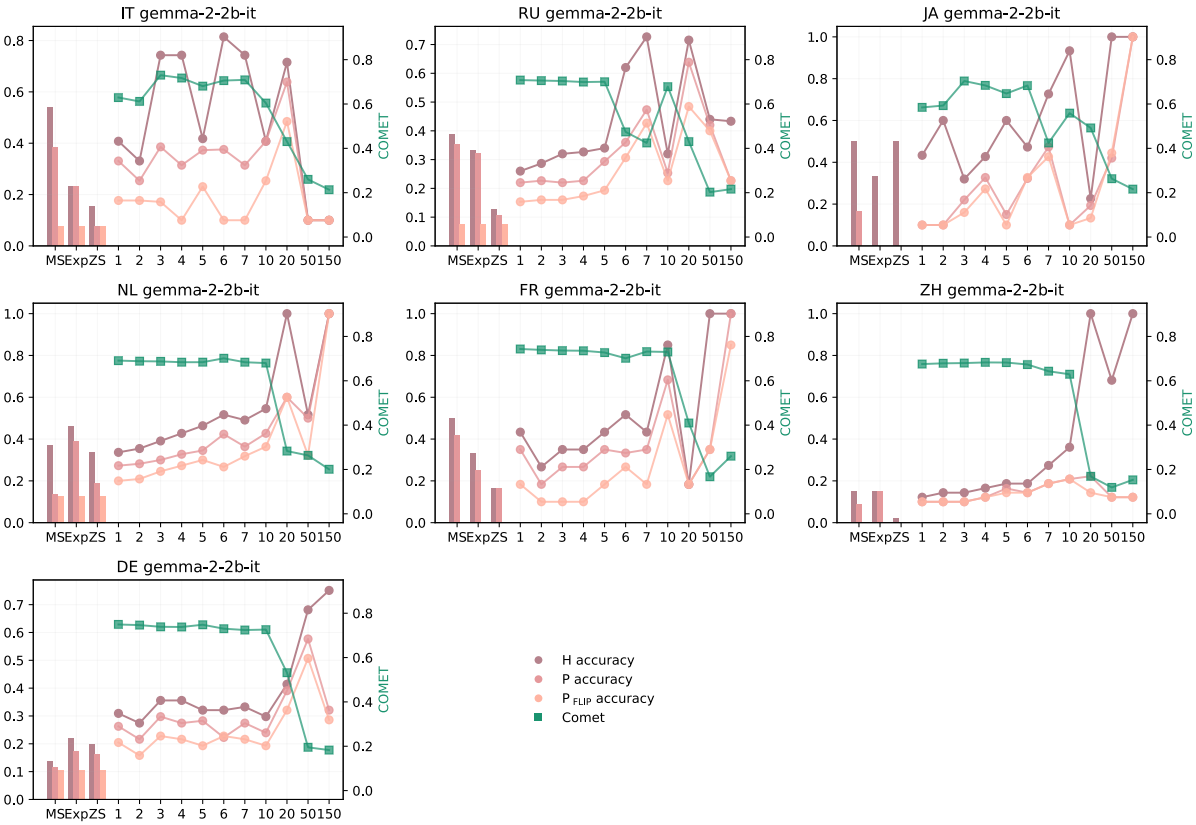

Figure 6: Results for every language on Gemma 2 2B.

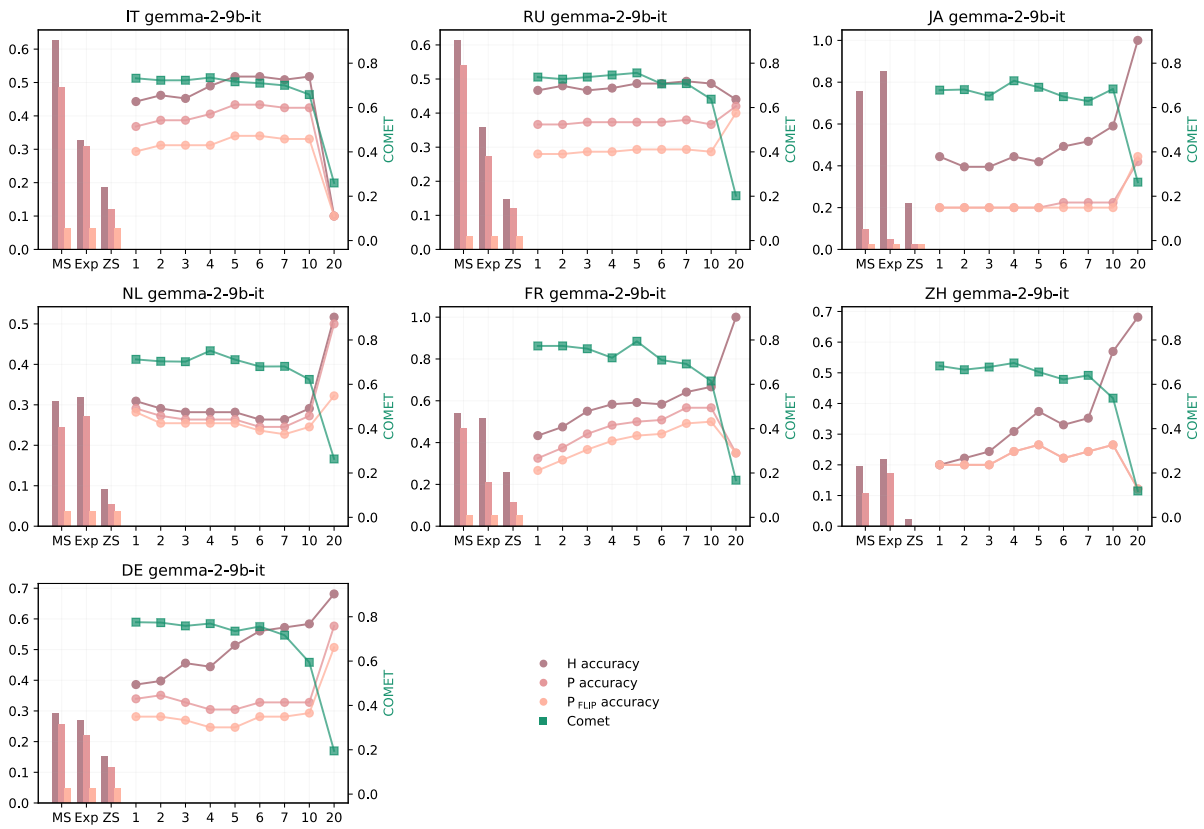

Figure 7: Results for every language on Gemma 2 9B.

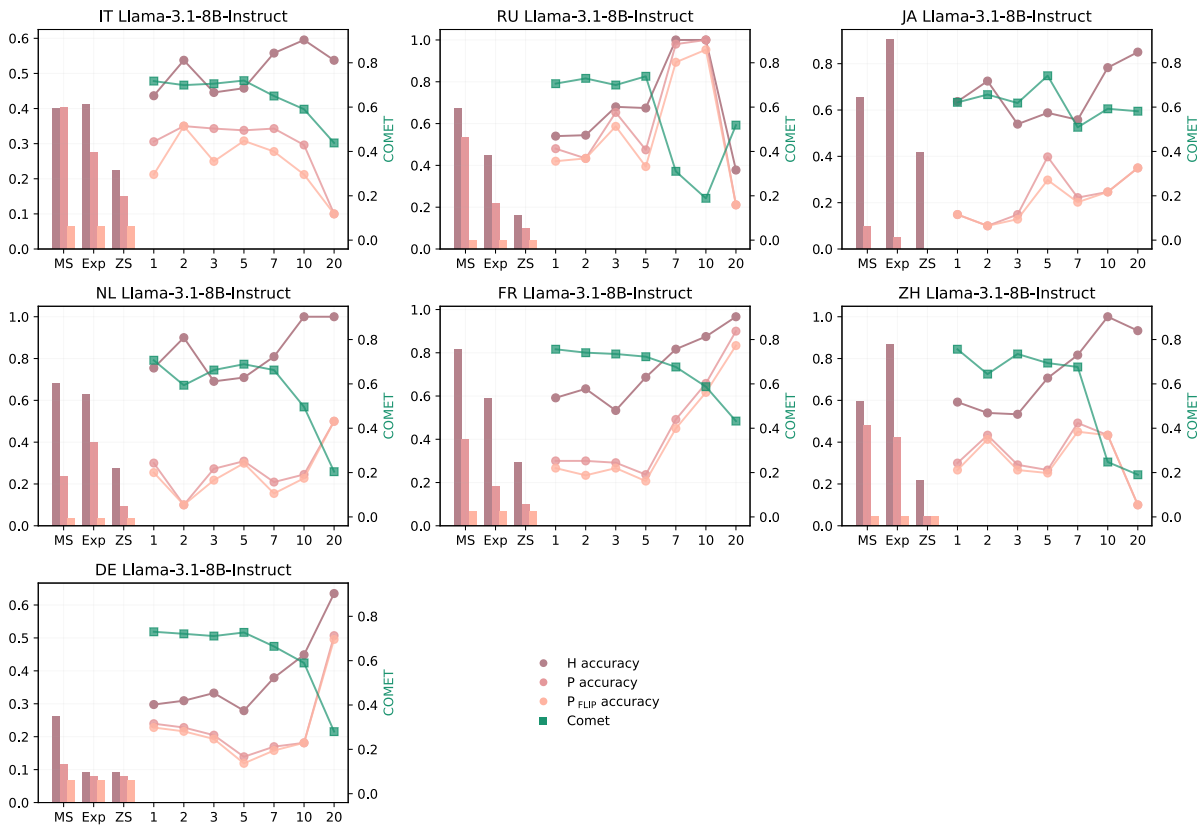

Figure 8: Results for every language on Llama 3.1 8B.

Source (ZH)	遂又往後看，只著一弓，弓上著香。也有一首歌云：「二十年辨是非，榴花照。三春及初春景，虎兕相逢大。」	classifier	Comet
H1	*Forthwith, therefore, he prosecuted a further perusal of what came next, when he caught sight of a picture of a bow. On this bow hung a citron. There was also this ode: Full twenty years right and wrong to expound will be thy fate! What place pomegranate blossoms come in bloom will face the Palace Gate! The third portion of spring, of the first spring in beauty short will fall! When tiger meets with hare thou wilt return to sleep perennial.*	H1	-
H2	*The picture that followed was of a bow with a citron hanging from it, followed by what looked like the words of a song: You shall, when twenty years in life's hard school are done, In pomegranate-time to palace halls ascend. Though three springs never could with your first spring compare, When hare meets tiger your great dream shall end.*	H2	-
Gemma 2 2B	*When he turned back to look, he saw a bow hanging on it, with a plum blossom. There was also a poem that said: "For twenty years, I have been trying to distinguish right from wrong, and the plum blossoms bloom in the palace. In the spring, the tigers and the rhinoceroses meet, and they dream of returning home."*	MT	0.49
Gemma 2 9B	*Looking back, he saw a painting of a bow with a quiver hanging on it. There was also a line of lyrics that said: Ëor twenty years, I have distinguished right from wrong, and the blossoming of the pomegranate flowers illuminates the imperial court. Three springs have passed, and now it is the beginning of spring again, like a dream where tigers and rhinoceroses meet.*	MT	0.56
Llama 3.1 8B	*As I looked back, I saw a painting of a bow with a censer on it. There was also a poem that said: T̈wenty years of distinguishing right from wrong, the lotus flower blooms in the palace. Three springs competing with the scenery of the first spring, the tiger and the rhinoceros meet and return to their dream."*	MT	0.53
ZS-Exp$_{M\leftrightarrow H}$	*And as he turned his gaze back upon the scene once more, there was displayed an archer's bow strung tautly; upon its surface hung a fragrant plum blossom. A poem also recited thus spake: "For twenty years I have sought truth amidst strife, where blooming durian flowers illuminate the palace walls. In spring's embrace, they vie for beauty, tigers clashing in dreams of return."*	MT	0.60
ZS-Exp$_{H_\alpha\leftrightarrow H_\beta}$	*And as he turned his gaze back upon the scene once more, there was displayed an archer's bow strung tautly; upon its surface hung a fragrant plum blossom. A poem also recited thus: "For twenty years I have sought truth amidst strife, where blooming durian flowers illuminate the palace walls. In spring's embrace, they vie for beauty, tigers clashing in dreams of return."*	MT	0.61
MS	*Then, turning around, he noticed a picture depicting a bow hanging on its string, adorned with incense sticks. A verse read: "For twenty years I've tried to discern truth from falsehood, and wherever fragrant plum blossoms bloom, the palace gates shine brightly."*	H1	0.65
ActAdd	*When he turned back to look, he saw a bow hanging on it, with a plum blossom. There was also a poem that said: "For twenty years, I have been trying to tell right from wrong, and the plum blossoms bloom in the palace. In spring, the tigers and rhinoceroses meet, and they dream of going back home."*	MT	0.53
ReFT	*When he turned back to look, he saw a bow hanging there, adorned with a plum blossom. There was also a poem that said: "For twenty years, I have tried to distinguish right from wrong, and plum blossoms bloom within the palace. In spring, tiger and rhinoceros meet, dreaming of their return home."*	MT	0.58
SAE Cont.$_{M\leftrightarrow H}$			
$\alpha = 1$	*When he turned back to look, he saw a bow hung upon it, with a plum blossom. There was also a verse which read: "For twenty years, I have struggled to discern truth from falsehood, and the plums blossomed where the imperial court resided. During the three springs, the tiger and rhinoceros met, dreaming of their return."*	MT	0.62
$\alpha = 5$	*He looked back and there hung a bow adorned with a plum blossom. A verse read: "Twenty years I' ve struggled for truth, where fragrant plums bloom near the imperial court. Spring arrives, tiger and rhino clash, dreaming of their return.*	MT	0.62
$\alpha = 150$	*The when the veryShellTruth lookWarmTheTheTheTigeratPoint [...]*	MT	0.11
SAE Cont.$_{H_\alpha\leftrightarrow H_\beta}$			
$\alpha = 5$	*He looked back and saw a bow adorned with fragrant citron. A verse declared: "Twenty years I' ve weighed right and wrong; where pomegranate blossoms bloom, they light the palace gates. Spring' s full bloom cannot match early spring' s grace. When tiger meets rhino, one returns to the great dream."*	H1	0.65

Table 8: Example paragraphs from "Dream of the Red Chamber" (ZH → EN) translated with H1personalization. Setup outputs are from Gemma 2 2B; only MS and SAE Cont.$_{H_\alpha\leftrightarrow H_\beta}$ successfully flip the classifier' prediction.

Source (IT)	Intanto la stalla fu chiusa e Pinocchio rimase solo: e perché erano molte ore che non aveva mangiato, cominciò a sbadigliare dal grande appetito. E, sbadigliando, spalancava una bocca che pareva un forno.	classifier	Comet
H1	*When the stable door was closed, Pinocchio was left alone. And since he hadn't eaten in many hours, he began to yawn from hunger. And when he yawned, his mouth opened as wide as an oven.*	H1	-
H2	*The stable was then shut and Pinocchio was left alone. He had not eaten for many hours, and he began to yawn from hunger. And when he yawned he opened a mouth that seemed as wide as an oven.*	H2	-
Gemma 2 2B	*At that moment the stable was closed and Pinocchio was left alone, and because it had been many hours since he had eaten, he began to yawn from a great hunger. And, yawning, he opened a mouth that seemed like a furnace.*	MT	0.79
Gemma 2 9B	*Meanwhile, the stable was closed and Pinocchio was left alone: and because it had been many hours since he had eaten, he began to yawn from his great hunger. And, yawning, he opened a mouth that looked like an oven.*	MT	0.80
Llama 3.1 8B	*Meanwhile the stable was closed and Pinocchio was left alone: and because it was many hours since he hadn't eaten, he began to yawn from great hunger. And, yawning, he opened a mouth that seemed like an oven.*	MT	0.80
ZS-Exp$_{M \leftrightarrow H}$	*Meanwhile, the stable door was shut, leaving Pinocchio alone. His tummy rumbled because it had been hours since his last meal, and he started yawning widely from his hunger. With each yawn, his mouth opened like an oven.*	MT	0.78
ZS-Exp$_{H_\alpha \leftrightarrow H_\beta}$	*Meanwhile, the stable door was shut, leaving Pinocchio all alone. His tummy rumbled because it had been hours since his last meal, and he started yawning widely from sheer hunger. With each yawn, his mouth gaped open like an oven.*	H2	0.80
MS	*Meanwhile, the stable door was shut and Pinocchio was left alone. Since it had been hours since he last ate, he began yawning from hunger. Each yawn opened his mouth wide enough to look like an oven.*	H2	0.81
ActAdd	*Meanwhile, the stable was closed and Pinocchio remained alone. Because many hours passed without eating, he started yawning from very big hunger. And while yawning, he opened his mouth wide like an oven.*	MT	0.75
ReFT	*Meanwhile, the stable was closed and Pinocchio stayed alone. Because many hours had passed since he last ate, he began to yawn from big hunger. And, yawning, he opened a mouth that looked like an oven.*	MT	0.77
SAE Cont.$_{M \leftrightarrow H}$			
$\alpha = 1$	*Meanwhile, the stable door was shut and Pinocchio was left alone. Because it had been many hours since he had eaten, he started to yawn with intense hunger. And, as he yawned, his mouth gaped open like an oven.*	MT	0.79
$\alpha = 5$	*Meanwhile, the stable door shut, leaving Pinocchio all alone. Because it had been many hours since he' d eaten, he started to yawn with intense hunger. As he yawned, his mouth gaped open, looking like an oven.*	H2	0.82
$\alpha = 50$	*He He "It The ... You It "[They They) THE...I I HE IT [There There THEYRE WE "A A FOR "[...]*	MT	0.14
SAE Cont.$_{H_\alpha \leftrightarrow H_\beta}$			
$\alpha = 5$	*The stable was then shut, leaving Pinocchio all alone. He had not eaten for many hours and began to yawn from great hunger. As he yawned, he opened his mouth wide, like an oven.*	H2	0.84

Table 9: Example paragraphs from "Pinocchio" (IT → EN) translated with H2personalization. Setup outputs are from Gemma 2 9B; ZS-Exp$_{H_\alpha \leftrightarrow H_\beta}$, MS, SAE Cont.$_{M \leftrightarrow H}$ and SAE Cont.$_{H_\alpha \leftrightarrow H_\beta}$ (both $\alpha = 5$) can flip the classifier prediction.

Source (FR)	Cette somme vous sera restituée á votre sortie de prison, dit le juge. En attendant, vous ê tes libres sous caution.	classifier	Comet
H1	*This sum of money will be returned to you when you leave prison,' said the judge. 'In the meantime you are free on bail.*	-	-
ZS	*This amount will be returned to you upon your release from prison, the judge said. Meanwhile, you are free on bail.*	MT	0.79
SAE Cont.$_{M \leftrightarrow H}$			
$\alpha = 5$	*This sum will be repaid to you at your departure from prison, the judge declared. In the meantime, you are released under bond.*	MT	0.85
$\alpha = 50$	*#echo all ought to goodness too was put moreover goodness behaving goodness goodness goodness goodness* [...]	H1	0.20
$\alpha = 150$	*GURL callers goodness_Noneummy amidála Reportedaps Beaverishlor Decompiled_unset queer headquarters Headquarters twe ... pocket-or callers dt behaving indoorsgetRepository callers_lazy uninsured tipostiFRING callers callers callers callers Guinness_None ... foss bedroomsitud facet goodness goodness goodness* [...]	H1	0.23
Source (DE)	Nein was Sie lieber haben, Herr Leutnant! Nur keine Zeremonien, es ist doch ganz einerlei.	classifier	Comet
H2	*"No, no-whatever you would rather have, Lieutenant Hofmiller! Please don't stand on ceremony, it makes no difference to us."*	-	-
ZS	*No, whatever you prefer, Lieutenant! Just no ceremonies, it doesn't matter.*	MT	0.76
SAE Cont.$_{M \leftrightarrow H}$			
$\alpha = 5$	*No, anything at all you want, sir! Just don't make a fuss about it, it really doesn't matter.*	H2	0.79
$\alpha = 50$	*">I Don't worry about that... I don't want a ceremony for this one. It's not important...*	H2	0.46
$\alpha = 150$	*IWhenInWhatItDonIf Sometimes AIs Celebrating cerimonies... Sosir please don't have parties ey'* [...]	H2	0.24

Table 10: Examples from different languages being classified as Human when using extreme α values.