Feature-Level Insights into Artificial Text Detection with Sparse Autoencoders

Kristian Kuznetsov[1,2], Laida Kushnareva[2], Polina Druzhinina[1, 5], Anton Razzhigaev[1, 5], Anastasia Voznyuk[3], Irina Piontkovskaya[2] Evgeny Burnaev[1, 5] Serguei Barannikov[1, 4]

[1]Skolkovo Institute of Science and Technology, [2]AI Foundation and Algorithm Lab, [3]Advacheck OÜ, Estonia, [4]CNRS, Université Paris Cité, France, [5]Artificial Intelligence Research Institute (AIRI)

Skoltech

Skolkovo Institute of Science and Technology

Motivation

The increasing realism of LLM-generated text poses a major challenge for **artificial text detection (ATD)**. Many current ATD methods lack interpretability and robustness, leaving the specific linguistic and structural features of machine-generated text underexplored. To address this, we use **Sparse Autoencoders (SAEs) as a source of interpretable and generalizable features** for understanding and detecting AI-generated text.

Methods

Feature Extraction with SAEs: we use sparse autoencoders to extract interpretable features from the residual stream of Gemma-2-2B. Given hidden activations $\mathbf{x} \in \mathbb{R}^d$, the SAE performs:

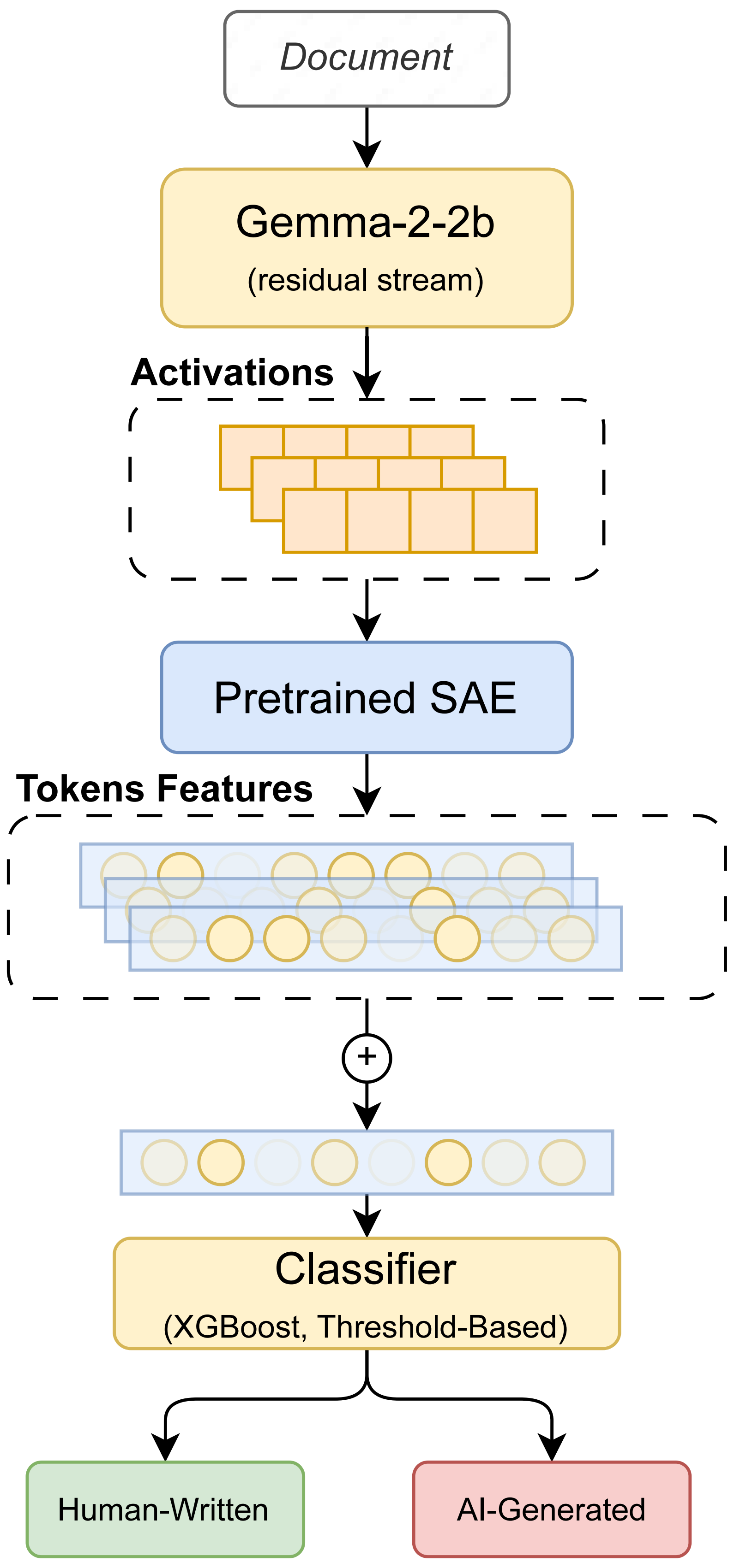
$$f(\mathbf{x}) = \sigma(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}), \quad \hat{x}(f) = \mathbf{W}_{\text{dec}} f(\mathbf{x}) + \mathbf{b}_{\text{dec}}$$

Here, $f(\mathbf{x}) \in \mathbb{R}^M$ (with $M \gg d$) is a sparse, non-negative feature vector. For sequence-level representation: $\mathbf{f} = \sum_{i=1}^{n} f^{(l)}(\mathbf{x}_i^{(l)})$, where $\mathbf{x}_i^{(l)}$ is the residual stream at token i in layer l.

Classification: we classify machine-generated and human-written texts from COLING dataset based on several stratagies: 1) **XGBoost** is trained on full feature vectors to measure global importance; 2) **Threshold classifiers** $\mathbb{I}[\mathbf{f}_j > \tau^*]$ and $\mathbb{I}[\mathbf{f}_j > 0]$ are used for analyzing individual feature activations.

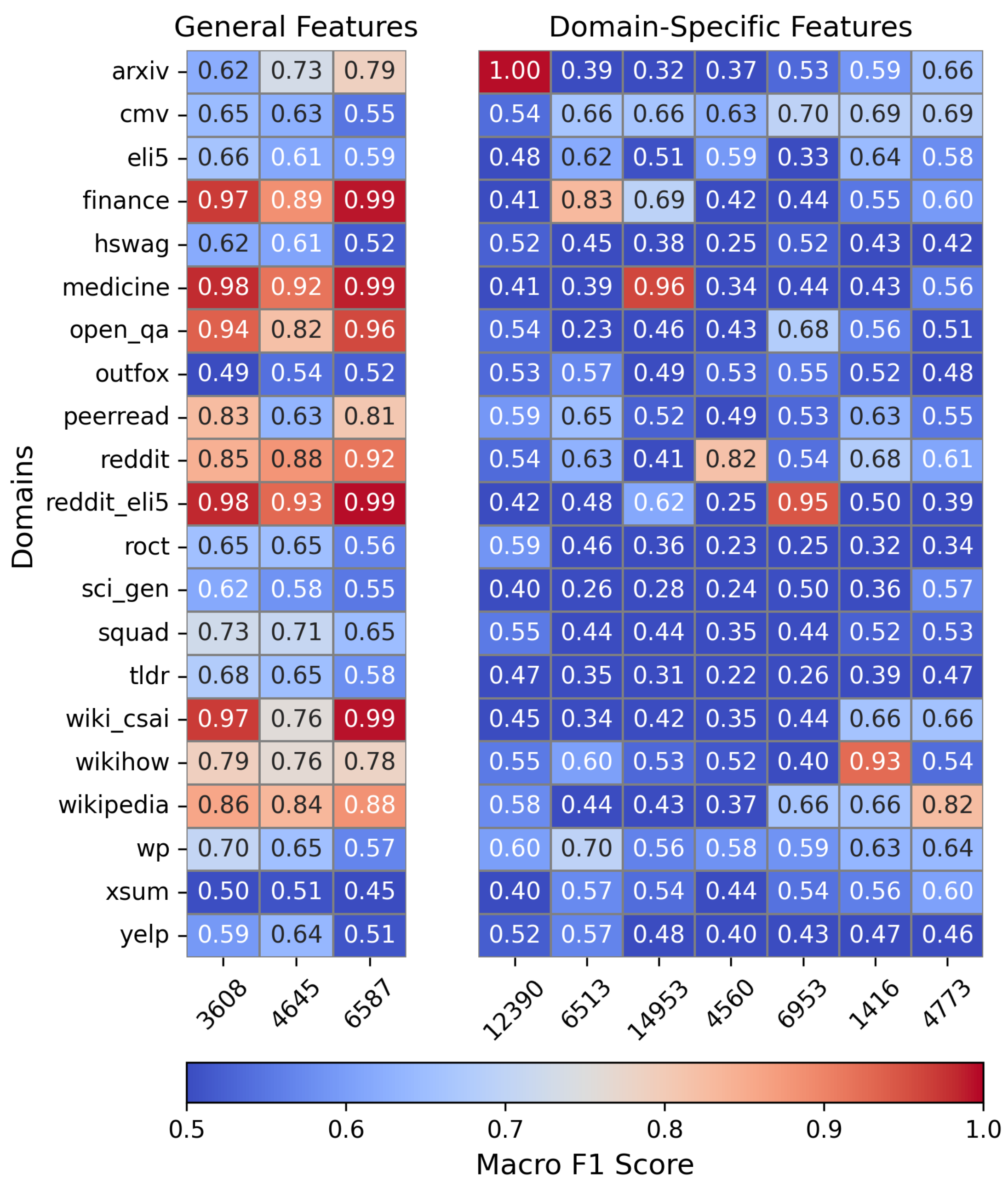
Feature Interpretation: we interpret the top-performing features using three complementary methods:

1. **Automatic Interpretation:** use GPT-4o to describe the semantic patterns in Top-N examples.
2. **Manual Inspection:** human analysis of Top-N examples to validate common linguistic traits.
3. **Steering + Interpretation:** Use feature steering $\mathbf{x}' = \mathbf{x} + \lambda A_{\max}\mathbf{d}_i$, where $\mathbf{d}_i$ is the i-th column of $\mathbf{W}_{\text{dec}}$, and analyze these generations for interpretability via GPT-4o.



Single Feature Classifiers

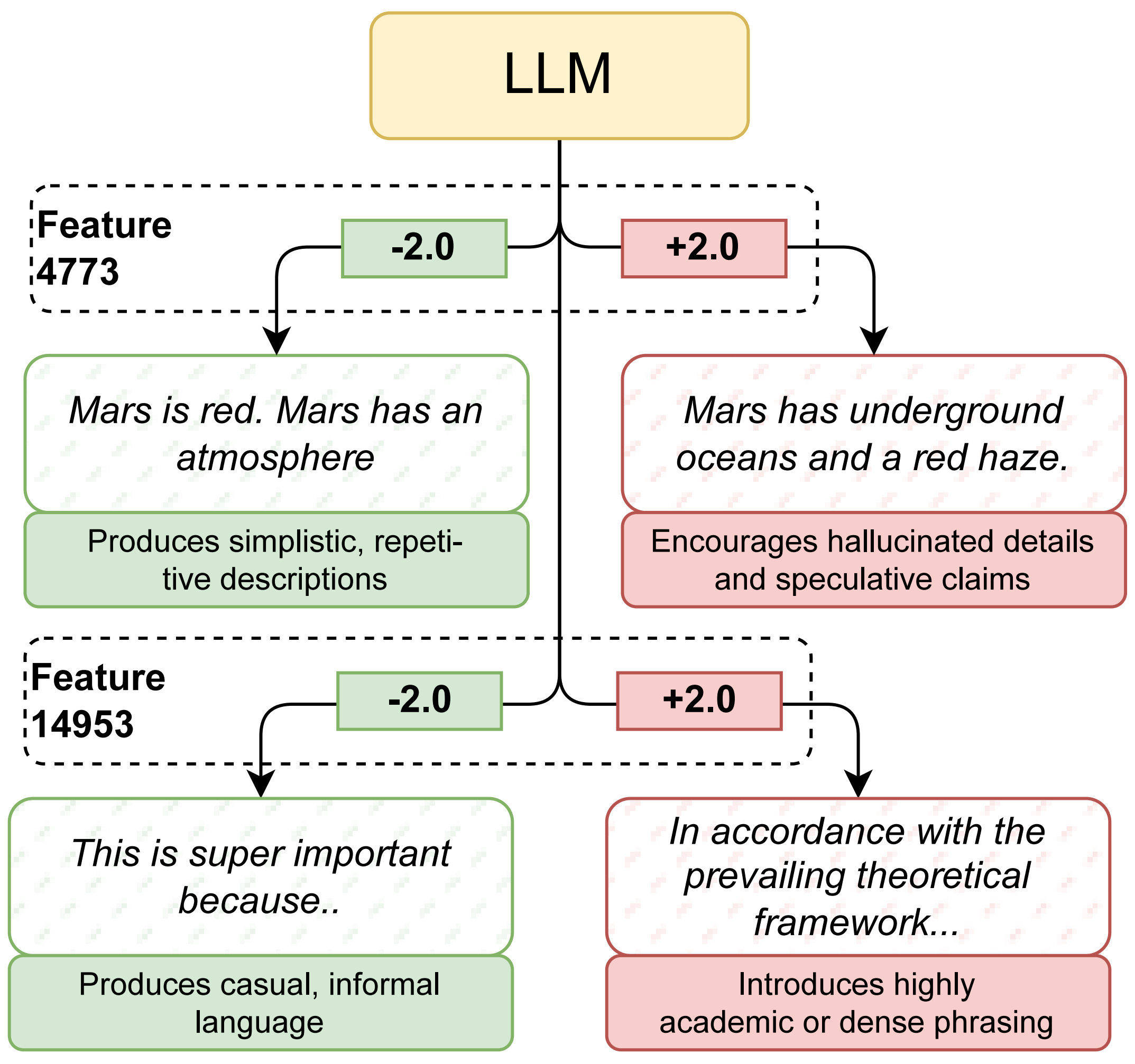
General Features: some features (e.g., 3608, 4645) generalize across domains and model families (e.g., Reddit, Wikipedia, Medicine).
Domain-Specific Features: others are highly domain-specific, capturing traits like hallucinated facts in Wikipedia.



Interpretation Insights

We identified several patterns in features and combined into groups:

1. **Common traits of AI-generated text**: *excessive complexity, assertive claims, wordy introductions, repetition, and overformality*
2. **Domain-specific signals**: *overcomplicated syntax, hallucinated facts, speculative tone, and overly polite phrasing depending on the source*



Mars is red. Mars has an atmosphere

Produces simplistic, repetitive descriptions

Mars has underground oceans and a red haze.

Encourages hallucinated details and speculative claims

This is super important because..

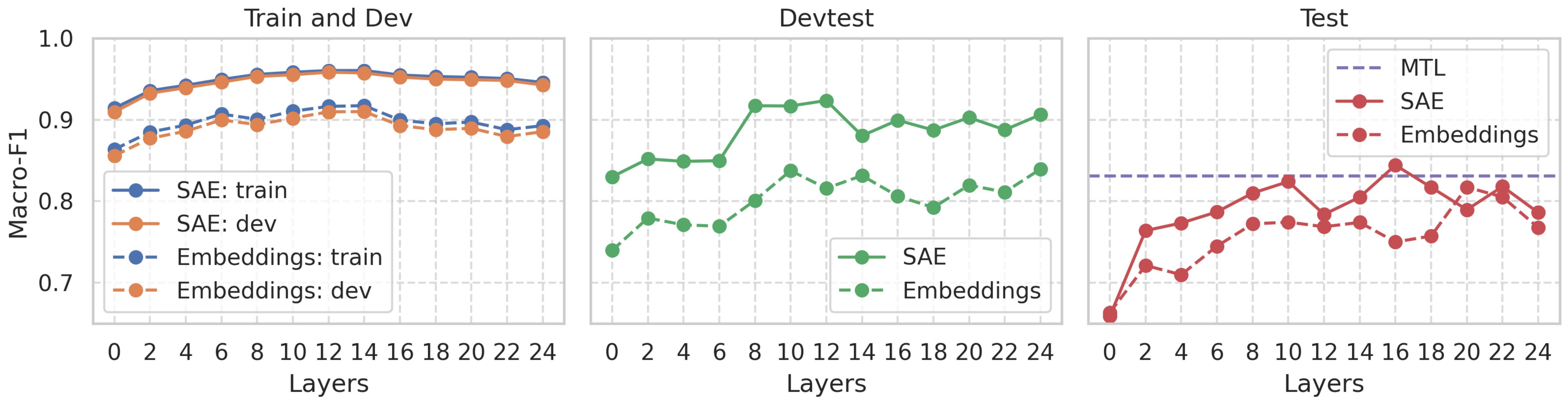
Produces casual, informal language

In accordance with the prevailing theoretical framework...

Introduces highly academic or dense phrasing

Overall Detection Performance

SAE-derived features consistently outperform mean-pooled transformer embeddings on the COLING dataset across all splits (Train, Dev, DevTest, Test). On the 16th layer, SAE-based classifiers surpass even the state-of-the-art multitask learning (MTL) baseline.



Conclusion

Sparse Autoencoders provide a powerful and interpretable alternative for detecting AI-generated text. A small number of learned features capture robust signals that **generalize across domains and models**. These features are not only effective for classification but also **human-interpretable through activation analysis, steering, and language-based explanations**. Our approach bridges performance and explainability, enabling more transparent and reliable AI-generated text detection.

Paper Link

For more information you can check paper via QR-code or contact me:

Telegram: @pyashy

Email: kris@kuznetsov.su