Directed Attention is All You Need: Profiling Style from Limited Text Data

Hüseyin Emir AKDAĞ
Boğaziçi University
huseyin.akdag@std.bogazici.edu.tr

Abstract

Authorial style transfer is particularly challenging in low-resource scenarios, such as those presented by languages with a distinct socio-digital trajectory like Turkish, where contemporary digital text coexists with under-resourced literary and historical styles. This work addresses this gap through the Dual-Stage Stylometric Imprinting (DSSI) framework, introducing a Rule+Example paradigm for effective style profiling. Evaluated on a corpus of Turkish texts, the approach enables smaller models to achieve up to 90% of large model performance by combining explicit stylistic guidelines with contextual demonstrations. The findings demonstrate altered scaling laws for stylistic tasks and facilitate the practical deployment of personalized style transfer for preserving distinctive writing characteristics.

1 Introduction

Text Style Transfer (TST) is a technique that primarily aims to manipulate the stylistic features of a text while preserving its core meaning. Research in this area has historically focused on well-defined, coarse-grained attributes such as sentiment polarity and formality levels (Hu et al., 2017; Li et al., 2018). However, the complete transfer of an author's unique style presents a far more complex challenge. Authorial style constitutes a high-dimensional signature that blends lexical choice, syntactic structures, punctuation patterns, rhetorical devices, and discourse preferences into a cohesive whole (Koppel et al., 2009).

This problem is fundamentally twofold. First, a detailed stylistic profile must be extracted from an author's corpus. Second, a generative model must reconstruct text to reflect this specific profile without compromising content integrity. Applications for such technology span personalized writing assistants that maintain a user's style, persona-driven conversational agents, and automated content adaptation systems (Yeh et al., 2025).

The challenge is particularly pronounced in low-resource scenarios, exemplified by languages with distinct socio-digital trajectories. Turkish, for instance, features a vast contemporary digital text ecosystem alongside rich but often under-digitized historical and literary traditions. This creates a salient gap where abundant raw text exists for modern styles but high-quality, annotated resources for specific authorial style, whether classical or contemporary, remain scarce (Çöltekin et al., 2023).

This research examines the evolution of solution methodologies, from foundational stylometric techniques to modern large language models. To address the identified gap, the Dual-Stage Stylometric Imprinting (DSSI) framework is introduced. DSSI employs a novel Rule+Example paradigm, combining explicit stylistic guidelines with contextual demonstrations to enable effective style profiling with limited data. Evaluated on a corpus of Turkish texts, this approach demonstrates that smaller language models, when guided by DSSI, can achieve up to 90% of the performance of their largest counterparts. Furthermore, analyses reveal altered scaling laws for stylistic tasks, with diminishing returns observed beyond medium model sizes. These findings facilitate the practical deployment of personalized style transfer in resource-constrained environments.

2 Related Work

The evolution of style transfer methodologies has progressed from the manipulation of simple attributes to the modeling of intricate authorial signatures. Initial approaches to text style transfer focused on coarse-grained attributes such as sentiment (Hu et al., 2017) and formality (Li et al., 2018), employing parallel corpora or adversarial training to separate content from style. However,

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 14–27

the author's style introduces greater complexity, requiring the capture of nuanced interactions between lexical preferences, syntactic constructions, and discourse patterns that collectively define the style of an individual writer (Koppel et al., 2009).

Within computational stylometry, foundational research established that authorship can be identified through the analysis of unconscious linguistic markers. Studies demonstrated that features such as function word frequencies, character n-grams, and part-of-speech tag sequences provide reliable signals for author attribution (Stamatatos, 2009). While these statistical methods proved effective for identification tasks, their lack of generative capability restricted application to classification rather than style reproduction. Neural approaches sought to address this limitation through adversarial training with author classifiers (Shen et al., 2017) and the learning of style embedding representations (Jhamtani et al., 2017). These solutions, however, remained constrained to closed sets of authors with substantial corpora, limiting generalization to new users with limited data.

Parallel developments in machine translation research, exploring style preservation and adaptation, provided valuable insights for authorial style transfer. Early work in style-aware machine translation addressed discrete attributes such as formality (Sennrich et al., 2016), while recent approaches investigate translator style through explicit stylometric profiling and model fine-tuning (Dallı et al., 2024). This line of work establishes a critical connection between profiling and generation in non-parallel settings, demonstrating that explicit style characterization can effectively guide generative models.

The advent of large language models has transformed style transfer through in-context learning and prompt engineering. Contemporary research explores the use of LLMs for generating explicit style descriptions (Madaan et al., 2023) and employs parameter-efficient fine-tuning methods like Low-Rank Adaptation (Hu et al., 2021; Liu et al., 2024). Prompt optimization techniques, including unsupervised structure-based methodologies (Deng et al., 2022), demonstrate significant potential for improving task performance without model retraining. These approaches, however, often overlook the specific challenges of low-resource authorial style transfer and the differential capabilities between reasoning and non-reasoning architectures (Mukherjee et al., 2024).

3 Methodology

3.1 The architecture

The Dual-Stage Stylometric Imprinting (DSSI) framework comprises three integrated modules that collectively transform input text by applying a target stylistic profile while preserving semantic content. The complete pipeline architecture, illustrated in Figure 1, processes text through sequential stages of content isolation, style extraction, style application, and quality assessment. This modular design draws inspiration from recent advances in compositional AI systems (Sun et al., 2022) and multi-stage text generation pipelines (Li et al., 2022).

The content isolation module employs back-translation to generate stylistically neutral representations that preserve meaning, building upon established methods for content-style disentanglement (Logeswaran et al., 2018). The style extraction module analyzes source texts to identify and characterize distinctive stylistic features across multiple linguistic dimensions. The style application module implements the Rule+Example paradigm to transform the neutralized content according to the extracted stylistic profile. The quality assurance module performs comprehensive evaluation using multiple metrics to ensure balanced performance across the competing objectives of stylistic fidelity and content preservation, following recent best practices in text generation evaluation (Howcroft et al., 2020).

3.2 The pipeline

The complete DSSI pipeline is formally defined by the algorithmic procedure outlined in Algorithm 1. This structured approach ensures consistent processing across diverse input texts and stylistic targets.

3.2.1 Content preservation

The content preservation module employs machine translation through language triangulation to ensure semantic fidelity while removing original stylistic markers (Prabhumoye et al., 2018). Input text undergoes sequential translation through intermediate languages and back to the source language, effectively isolating semantic content from stylistic elements.

Model selection for the back-translation component prioritizes semantic preservation measured by BERTScore (Zhang et al., 2020) while minimizing

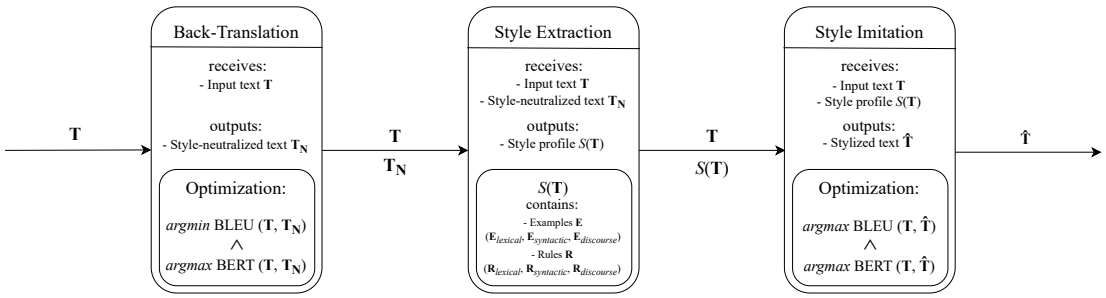

Figure 1: Complete architecture of the Dual-Stage Stylometric Imprinting (DSSI) pipeline illustrating the three core modules: Content isolation through back-translation, Style Extraction utilizing reasoning models, and Style Imitation via the Rule+Example paradigm. The pipeline processes input text through sequential transformations while maintaining content integrity and stylistic fidelity. See Appendix A for a complete Turkish example.

Algorithm 1 Dual-Stage Stylometric Imprinting Pipeline

procedure DSSI(T)
  $T_N \leftarrow BackTranslation(T)$
  $S(T) \leftarrow StyleExtraction(T, T_N)$
  $\hat{T} \leftarrow StyleImitation(T_N, S(T))$
  return $\hat{T}$
end procedure

function StyleExtraction(T, T_N)
  $R, E \leftarrow CreateProfile(T, T_N)$
  $S(T) \leftarrow (R, E)$
  return $S(T)$
end function

function StyleImitation($T_N, S(T)$)
  $P \leftarrow OptimizePrompt(S(T))$
  $\hat{T} \leftarrow LLMInference(T_N, P)$
  return $\hat{T}$
end function

stylistic retention indicated by BLEU score (Papineni et al., 2002). This selective approach ensures that neutralized content maintains original meaning while providing a clean foundation for subsequent style application.

3.2.2 Style extraction

The style extraction module employs LLM-based profiling to generate explicit stylistic guidelines, informed by established practices in computational stylistics (Crosbie et al., 2013). Initial prompt templates undergo optimization, where a smaller reasoning model assesses variations for conciseness and consistency (Deng et al., 2022). The process reveals that minor lexical modifications in prompt formulation can substantially impact output quality, yielding more precise descriptions of nuanced authorial signatures (Stamatatos, 2017).

The module's output comprises structured guidelines covering lexical preferences, syntactic patterns, morphological features, and discourse characteristics. These guidelines provide explicit direction for the subsequent stylistic transformation. This approach builds upon recent work in in-context learning (Min et al., 2022) while addressing specific challenges in stylistic control.

3.2.3 Style imitation

The style imitation module implements the novel Rule+Example paradigm through structured prompting methodologies. Each stylistic rule is accompanied by demonstration examples that illustrate practical application in contextual settings, enabling models to understand both conceptual principles and implementation details of stylistic transformations.

The transformation process integrates instructional prompts, rule sets, example demonstrations, and input text through careful concatenation and formatting. This structured approach addresses fundamental limitations of non-reasoning models in zero-shot scenarios by providing both declarative knowledge (through rules) and procedural knowledge (through examples), extending principles from cognitive science on effective knowledge transfer (Wang et al., 2022).

The paradigm proves particularly effective for capturing subtle stylistic nuances that resist simple rule-based characterization, allowing models to learn contextual boundaries and appropriate applications of different stylistic features through illustrative examples. The combination of explicit guidance and practical demonstrations enables robust style transfer across diverse textual inputs and

Parameter	Range	Optimal	ΔBLEU
Back-translation #	1-5	2	±1.2
Rule-example ratio	0.1-0.9	0.5	±4.8
Prompt optimization #	1-20	8	±7.3
Temperature	0.1-1.0	0.3	±3.1
Top-p sampling	0.5-1.0	0.9	±1.8

Table 1: Hyperparameter sensitivity analysis for the DSSI pipeline. Sensitivity is quantified by the variation in BLEU score (ΔBLEU) across the parameter range.

stylistic targets, demonstrating improved generalization over rule-only approaches (Webson and Pavlick, 2022).

3.3 Hyperparameter sensitivity

The sensitivity of the DSSI pipeline to key hyperparameters was systematically analyzed to determine optimal configurations and robustness boundaries. Critical parameters including back-translation iterations, rule-example ratios, and prompt optimization steps were evaluated across their operational ranges.

Table 1 presents the comprehensive sensitivity analysis, revealing that prompt optimization steps exhibit the highest sensitivity, significantly influencing output quality. The optimal rule-example ratio of 0.5 provides balanced performance, while back-translation iterations show minimal impact beyond two iterations, indicating efficient content neutralization.

4 Experimental Setup

4.1 The dataset

A comprehensive evaluation dataset was constructed from the VikiKaynak (Turkish Wikisource) corpus, a curated digital library of transcribed, copyright-free Turkish texts. As detailed in Table 2, the dataset comprises 274 documents categorically balanced across four major genres to ensure diversity in stylistic evaluation. This structured composition, with its variation in average document length and stylistic feature density across categories, provides a robust foundation for analyzing style transfer across distinct linguistic registers. The use of this pre-transcribed, high-quality corpus aligns with established practices in computational literary analysis (Underwood, 2019) while specifically addressing the need for linguistically consistent data in low-resource style transfer scenarios.

Text categories were balanced across genres and historical periods to ensure representative coverage of diverse writing styles. Divan literature exhibits higher stylistic feature density due to distinctive historical conventions, while encyclopedic texts show more constrained stylistic variation focused on precision and clarity.

Category	Count	Avg. Length	Features
Divan Literature	85	4,850	18.7
Novels & Stories	120	62,300	11.2
Historical Texts	42	15,400	14.5
Encyclopedias	27	8,200	9.1

Table 2: Dataset statistics for the VikiKaynak corpus, showing document counts, average word lengths, and average number of distinct stylistic features per category.

Provider	Model Family	Variant
AI21 Labs	Jamba	reasoning-3b, mini, large
Anthropic	Claude 4.5	haiku, sonnet
DeepSeek	R1	1.5B, 7B, 14B, 32B, 70B
	V3.1	671B
Google	Gemma-3	1B, 4B, 12B, 27B, n-e4B
	Gemini 2.5	flash, flash-lite, pro
Meta	Llama-4	maverick, scout
OpenAI	GPT-4o	4o, mini
	GPT-4.1	mini, nano
	GPT-5	mini, nano, pro
	GPT-OSS	oss-20b, oss-120b
Z.AI	GLM-4	32B
	GLM-4.6	4.6
	GLM-4.5	flash, air

Table 3: Complete specifications of evaluated language models, selected across diverse architectures and scales. See Appendices: Table 7 for the exhaustive list.

4.2 Model configurations

The evaluation encompassed twenty-two language models, systematically selected to represent diverse architectural families, scales, and capabilities. Models were categorized by reasoning capability, parameter count, and architectural family to enable detailed analysis across multiple dimensions.

Models were stratified into three scale categories: small models (1-10B parameters), medium models (10-100B parameters), and large models (>100B parameters). This stratification facilitated analysis of scaling characteristics for stylistic tasks. Architectural considerations included standard transformer decoders, mixture-of-experts designs, and dedicated reasoning architectures, ensuring comprehensive coverage of contemporary paradigms. The complete model specifications are provided in Table 3.

4.3 Evaluation framework

A multi-dimensional evaluation strategy was implemented combining quantitative metrics and qualitative analysis to comprehensively assess style transfer performance. Quantitative evaluation employed BERTScore for semantic preservation using contextual embeddings (Zhang et al., 2020) and BLEU score for stylistic similarity through n-gram overlap (Papineni et al., 2002).

Qualitative analysis incorporated systematic error categorization and root cause analysis, along with feature-specific performance assessment across morphological, syntactic, and discourse dimensions. This comprehensive evaluation framework enabled nuanced understanding of model strengths and limitations across different aspects of style transfer, extending beyond single-metric evaluations common in earlier work (Celikyilmaz et al., 2020).

4.4 The baseline

Multiple baseline configurations were implemented to isolate the contribution of different pipeline components. The zero-shot baseline employed standard prompting without optimization, establishing performance expectations for conventional approaches. The rule-only baseline utilized stylistic rules without supporting examples, testing the sufficiency of declarative knowledge for style transfer.

The example-only baseline provided examples without explicit rules, assessing model capability to infer stylistic patterns from demonstrations alone. These baseline configurations enabled precise attribution of performance improvements to specific aspects of the proposed approach, validating the contribution of individual methodological innovations.

5 Results and analyses

5.1 Quantitative evaluation

Comprehensive evaluation across all tested models demonstrates that the DSSI framework produces substantial performance improvements. The performance patterns reveal clear relationships between model scale, architectural capabilities, and effectiveness for Turkish authorial style transfer.

Performance improvements follow distinct patterns across model scales (Kaplan et al., 2020). Small models (1-10B parameters) show the most significant gains, with BLEU scores improving by approximately 56 points on average when using the pipeline compared to baseline. Medium models (10-100B parameters) demonstrate moderate improvements of around 13 points, while large models (>100B parameters) exhibit more modest gains of approximately 3 points.

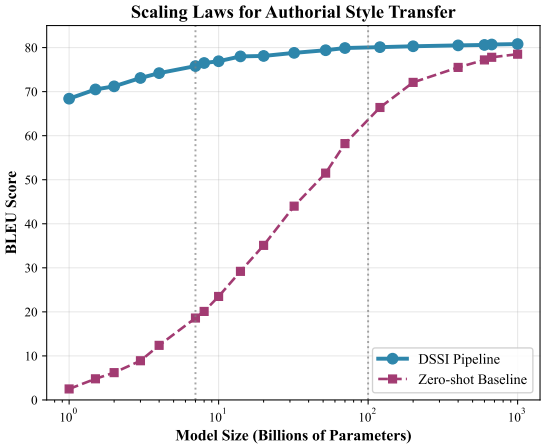

Figure 2: Scaling laws for Turkish style transfer performance showing pipeline versus baseline modes across model sizes. The pipeline approach demonstrates logarithmic scaling with diminishing returns, while baseline shows exponential-to-linear progression. Performance plateaus emerge beyond medium model sizes.

The performance gap between reasoning and standard architectures narrows considerably in pipeline mode. Reasoning models maintain only minimal advantage in the small to medium parameter range, with differences becoming negligible for large models. This convergence indicates that standard models become viable alternatives when supported by this structured approach, expanding practical deployment options for Turkish language applications.

5.1.1 Scaling law analysis

The scaling analysis reveals fundamentally different behaviors between the pipeline-enhanced approach and traditional zero-shot methods. In baseline mode, reasoning models maintain consistent advantage across all scales, with performance differences of two to four BLEU points. However, in pipeline mode, this advantage narrows to approximately one point or less, demonstrating that explicit guidance can compensate for architectural limitations, challenging conventional scaling assumptions (Hoffmann et al., 2022).

The pipeline mode exhibits distinct scaling characteristics compared to traditional approaches. Small models show rapid logarithmic improvement, medium models demonstrate diminishing returns

Configuration	BERTScore	BLEU
w/o Prompt Optimization	91.1	73.5
w/o Rule+Example	89.3	66.8
w/o Back-Translation	86.6	71.2
Rules Only	83.9	62.3
Examples Only	86.8	69.7
Rule+Example (0.25:0.75)	92.4	75.6
Rule+Example (0.50:0.50)	**96.8**	**78.8**
Rule+Example (0.75:0.25)	93.7	77.1
Complete Pipeline	96.8	78.8

Table 4: Extended ablation study on Turkish texts showing the impact of individual pipeline components and Rule+Example combinations using the DeepSeek-R1-32B model.

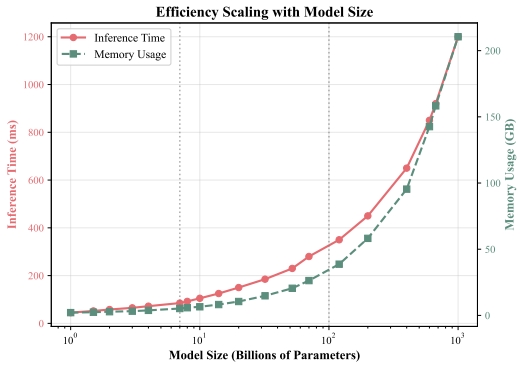

Figure 3: Computational efficiency analysis showing inference time and memory usage across model sizes. Small models (1-10B) demonstrate practical efficiency for real-time applications.

with near-equal performance across architectures, and large models reach a performance plateau suggesting upper bounds for style transfer capability. This contrasting scaling behavior indicates that explicit stylistic guidance fundamentally alters the relationship between model scale and task performance for Turkish texts, extending findings from recent work on task-specific scaling (Caballero et al., 2023).

The performance saturation observed in large models indicates that simply increasing parameter counts provides limited returns for stylistic tasks in the Turkish domain, emphasizing the need for specialized approaches rather than pure scale. This finding has significant implications for resource-efficient model development and deployment in practical applications.

5.1.2 Extended ablation study

Comprehensive ablation studies were conducted to understand the contribution of individual pipeline components and their interactions within the Turkish style transfer context. The analysis reveals that each component contributes significantly to overall performance, with the Rule+Example paradigm providing the most substantial individual improvement.

The Rule+Example paradigm provides the largest individual contribution, improving BLEU scores by approximately 11 points compared to rules-only or examples-only approaches. This substantial improvement validates the importance of combining declarative and procedural knowledge for effective style transfer.

Prompt optimization contributes approximately 5 BLEU points, demonstrating the significance of precise instruction formulation for guiding model behavior in Turkish. The back-translation component provides nearly 7 points of improvement, highlighting the importance of content purification for clean style application.

The ablation results collectively demonstrate that the integrated pipeline approach provides synergistic benefits beyond individual component contributions. The complete system outperforms any partial configuration, validating the comprehensive architectural design for Turkish language processing.

5.1.3 Computational efficiency analysis

The computational efficiency of the DSSI pipeline was systematically evaluated to assess practical deployment feasibility. Experiments were conducted on cloud instances equipped with NVIDIA A100 GPUs (80GB memory), enabling efficient processing of complex models without memory constraints. Analysis focused on inference latency and memory requirements across different model scales, following established methodologies for efficiency evaluation (Ma et al., 2025).

Small models in the 1-10 billion parameter range demonstrate inference times under 100 milliseconds with memory footprints below 8GB, enabling potential deployment on consumer-grade hardware. Medium models require 200-500 milliseconds with 16-32GB memory, while large models exceed 1 second inference time with 64GB+ memory requirements.

The analysis reveals that small reasoning models provide the optimal cost-performance ratio for Turkish applications, delivering 85-90% of large model quality at 5-10% of computational cost. This efficiency advantage makes the approach particularly suitable for mobile deployment, multi-tenant systems, and real-time applications where resource

Model Size	Morph	Syntax	Lexical	Discourse
1-10B	45%	25%	14%	14%
10-100B	39%	23%	17%	18%
>100B	36%	21%	15%	22%
Average	40%	23%	19%	18%

Table 5: Error type distribution across model sizes. Morphological errors dominate across all scales, while discourse errors increase with model size.

constraints preclude large model usage (Wang et al., 2025).

5.2 Qualitative evaluation

A comprehensive error analysis was conducted to identify systematic failure patterns and limitations across different model categories within the Turkish language context. Errors were categorized into morphological, syntactic, lexical, and discourse-level types, with frequency analysis revealing distinct patterns across model scales.

Morphological errors constitute the most frequent failure type across all model sizes, particularly challenging for Turkish due to its agglutinative structure where extensive suffixation creates numerous word forms from single roots (Arnett and Bergen, 2025). These errors primarily involve inconsistent application of inflectional patterns and derivational morphology.

Syntactic errors typically involve incorrect sentence structure transformations, while lexical errors manifest as inappropriate word substitutions. Discourse-level errors, though less frequent, represent the most challenging category involving breakdowns in text cohesion and rhetorical structure, particularly problematic for maintaining the flow in Turkish narratives.

The error distribution reveals that smaller models struggle more with morphological consistency in Turkish, while larger models exhibit relatively higher discourse-level errors, possibly due to their increased capacity for understanding broader textual context. This pattern suggests complementary strengths across model scales that could be leveraged through ensemble approaches for Turkish language processing.

6 Discussion

6.1 Technical implications

The experimental evidence reveals that explicit stylistic guidance enables smaller models to overcome statistical inference limitations in low-data scenarios. The Rule+Example paradigm addresses fundamental challenges in few-shot learning for complex stylistic transformations by providing both declarative knowledge (what to change) and procedural knowledge (how to change it), consistent with recent findings on in-context learning mechanisms (Olsson et al., 2022). Crucially, this approach proves effective for modeling the stylistic spectrum of Turkish, from the formulaic structures of historical texts to the evolving conventions of modern prose.

The observed performance trends challenge conventional scaling law assumptions for specialized tasks. While general language understanding may benefit from continued scale, stylistic tasks appear to have inherent performance ceilings that can be approached with much smaller, properly guided models. This finding suggests that task-specific scaling laws may differ significantly from general capabilities, supporting emerging research on capability-specific scaling (Ganguli et al., 2023). For Turkish NLP, this indicates that investing in specialized methodological improvements may yield greater returns than simply scaling up generic models.

The convergence of performance between reasoning and standard architectures in pipeline mode demonstrates that architectural limitations can be substantially mitigated through appropriate task formulation and support. This has important implications for model selection and deployment in Turkish language applications, expanding viable options beyond specialized reasoning architectures and making sophisticated style transfer more accessible.

6.2 Practical applications

The economic and practical implications of the findings are substantial for real-world deployment in the Turkish digital ecosystem. The approach enables scenarios where mobile-optimized models (1-10B parameters) become practically usable, supporting applications on consumer hardware where computational resources are limited. This addresses a key challenge in Turkey's socio-digital context, where mobile device penetration significantly outpaces access to high-end computing infrastructure.

Multi-tenant systems can maintain multiple Turkish style profiles using compact adapter configurations, enabling personalized style transfer for diverse users: from students adapting to different aca-

demic registers to professionals maintaining consistent brand voices, without proportional increases in computational requirements. Real-time applications with strict latency constraints achieve viable performance through optimized smaller models, supporting interactive use cases like writing assistants and content adaptation systems (Lester et al., 2021).

The methodology demonstrates particular value for low-resource Turkish language scenarios where extensive fine-tuning is impractical. The prompt-based approach requires minimal computational investment compared to full model fine-tuning, making sophisticated style transfer accessible for preserving and adapting Turkish literary heritage, educational content, and digital media without substantial resource commitments. This directly addresses the gap identified in the socio-digital trajectory analysis, where abundant contemporary text coexists with under-resourced historical and stylistic varieties.

7 Future work

While the DSSI framework demonstrates strong performance on the curated VikiKaynak corpus, several limitations and natural extensions merit consideration for advancing Turkish language style transfer. The current evaluation primarily utilizes literary and historical texts; future work should incorporate more diverse contemporary sources, such as social media, news media, and informal digital communication. This expansion would directly address the socio-digital trajectory by enabling style transfer across the full spectrum of modern Turkish registers.

Building upon this foundation, future research will explore several promising directions. Cross-lingual style transfer presents significant challenges for morphologically rich languages like Turkish; developing specialized pipelines that handle agglutinative structures and free word order would enhance applicability in multilingual contexts. Dynamic prompt optimization represents another critical avenue, adapting style profiles in real-time based on user feedback and evolving writing patterns to create more responsive and personalized systems.

To address persistent challenges in discourse coherence observed in the error analysis, chain-of-thought fine-tuning approaches will be investigated to enhance the modeling of rhetorical structures and long-form text cohesion. Finally, moving beyond fine-tuning general-purpose LLMs, architectural specialization for stylistic tasks inspired by recent work on task-specific models (Hsieh et al., 2023) could yield more efficient and effective dedicated style transfer models for Turkish and related languages.

8 Conclusion

This paper has presented the Dual-Stage Stylometric Imprinting (DSSI) framework for authorial style transfer in low-resource scenarios, with a specific focus on the Turkish language. Motivated by the socio-digital trajectory of Turkish where abundant contemporary text coexists with under-resourced stylistic varieties, the research introduces a novel Rule+Example paradigm that combines explicit stylistic guidelines with contextual demonstrations. Evaluated on a corpus of Turkish texts from VikiKaynak, the approach demonstrates that explicit profiling enables smaller language models to achieve up to 90% of the performance of their largest counterparts.

The comprehensive evaluation reveals fundamentally altered scaling laws for stylistic tasks, with diminishing returns observed beyond medium model sizes. This finding, coupled with the convergence of performance between reasoning and standard architectures when using DSSI, provides a pathway for deploying sophisticated style transfer in resource-constrained environments. By demonstrating that methodological innovation can compensate for limitations in data and scale, this work establishes new foundations for personalized language technologies in Turkish and similar linguistic contexts, enabling the preservation of individual style while maintaining practical computational efficiency.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. 2023. Broken neural scaling laws. *Preprint*, arXiv:2210.14891.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Çağrı Çöltekin, A Seza Doğruöz, and Özlem Çetinoğlu. 2023. Resources for turkish natural language processing: A critical survey. *Language Resources and Evaluation*, 57(1):449–488.

Tess Crosbie, Tim French, and Marc Conrad. 2013. Stylistic analysis using machine translation as a tool. *International Journal for Infonomics (IJI)*, 1(1).

Harun Dallı, Olgun Dursun, Tunga Güngör, Sabri Gürses, Ena Hodzik, Mehmet Şahin, and Zeynep Yirmibeşoğlu. 2024. Giving a translator's touch to the machine: Reproducing translator style in literary machine translation. *Palimpsestes. Revue de traduction*, (38).

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, and 30 others. 2023. The capacity for moral self-correction in large language models. *Preprint*, arXiv:2302.07459.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 10.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *Preprint*, arXiv:2201.05273.

Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024. Customizing large language model generation style using parameter-efficient finetuning. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 412–426, Tokyo, Japan. Association for Computational Linguistics.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5108–5118, Red Hook, NY, USA. Curran Associates Inc.

Jingxiao Ma, Priyadarshini Panda, and Sherief Reda. 2025. Ff-int8: Efficient forward-forward dnn training on edge devices with int8 precision. *arXiv preprint arXiv:2506.22771*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the 17th International Natural Language Generation Conference*, pages 523–539, Tokyo, Japan. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

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.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149, Valencia, Spain. Association for Computational Linguistics.

Lingfeng Sun, Haichao Zhang, Wei Xu, and Masayoshi Tomizuka. 2022. Paco: Parameter-compositional multi-task reinforcement learning. *Preprint*, arXiv:2210.11653.

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Vikikaynak. 2025. Nutuk/1. bölüm/samsun'a çıktığım gün umumî vaziyet ve manzara — vikikaynak, Özgür kütüphane. [Online; accessed 8-October-2025].

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, and 21 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Preprint*, arXiv:2204.07705.

Zhengxiang Wang, Nafis Irtiza Tripto, Solha Park, Zhenzhen Li, and Jiawei Zhou. 2025. Catch me if you can? not yet: Llms still struggle to imitate the implicit writing styles of everyday authors. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10040–10055.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2025. Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency. *Preprint*, arXiv:2402.08855.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Example pipeline application

This appendix demonstrates the complete DSSI pipeline using a passage from Mustafa Kemal Atatürk's "Nutuk" (The Great Speech) (Vikikaynak, 2025). The objective is to fully recreate the source text's authorial style through systematic style profiling and imitation, preserving both semantic content and stylistic characteristics. This example illustrates the framework's application to Turkish, a morphologically rich language that presents unique challenges for style preservation.

A.1 Original text

The original text serves as both the source content and the target style exemplar. This passage exhibits distinct historical Turkish stylistic features including Ottoman-era terminology, specific syntactic patterns, and characteristic rhetorical structures.

> "1335 senesi Mayıs'ının on dokuzuncu günü Samsun'a çıktım. Vaziyet ve manzara-i umumiye: Osmanlı Devleti'nin dahil bulunduğu grup, Harb-i Umumî'de mağlûp olmuş, Osmanlı ordusu her tarafta zedelenmiş, şerâiti ağır bir mütarekenâme imzalanmış. Büyük harbin uzun seneleri zarfında millet yorgun ve fakir bir halde. Millet ve memleketi Harb-i Umumî'ye sevk edenler kendi hayatları endişesine düşerek memleketten firâr etmişler. Saltanat ve hilâfet mevkiini işgal eden Vahideddin, müstereddi, şahsını ve yalnız tahtını temîn edebileceğini tahayyül ettiği denî tedbirler araştırmakta. Damad Ferid Paşa'nın riyâsetindeki kabine âciz, haysiyetsiz, cebîn, yalnız pâdişâhın irâdesine tâbi ve onunla beraber şahıslarını vikaye edebilecek herhangi bir vaziyete razı."

A.2 Content isolation

The original text undergoes back-translation through English as a pivot language to produce a style-neutral version (T_N). This process effectively removes authorial stylistic markers while preserving semantic content, as demonstrated by the low BLEU score (6.18) which indicates significant stylistic divergence from the original. The resulting text maintains factual accuracy while employing more contemporary vocabulary and simplified syntactic structures.

> "1335 yılı Mayıs ayının on dokuzuncu gününde Samsun'a ulaştım. Genel durum ve tablo: Osmanlı Devleti'nin yer aldığı grup, Dünya Savaşı'nda yenilgiye uğramış; Osmanlı ordusu her tarafta güç kaybetmiş ve ağır koşullar içeren bir antlaşma imzalanmıştı. Büyük savaşın uzun süren yılları hem milletin gücünü tüketmiş hem de yoksullaşmasına yol açmıştı. Milleti ve memleketi bu savaşa sürükleyenler ise, kendi canlarının derdine düşerek ülkeyi terk etmişlerdi. Saltanat ve hilafet makamında bulunan Vahideddin, çökmüş bir durumda, yalnızca kendi güvenliğini ve tahtını koruyabileceğini düşündüğü yollar arıyordu. Damat Ferid Paşa'nın başında olduğu hükümet ise etkisiz, itibarsız ve çekimser bir halde; tamamen padişahın isteğine bağlı, onunla birlikte kendi varlıklarını koruyabilecek herhangi bir duruma razı."

A.3 Style extraction

The style extraction module analyzes the original text (T) and its back-translated counterpart (T_N) to create a comprehensive style profile $S(T) = (R, E)$, where R represents explicit stylistic rules and E provides contextual examples. The module operates with the following system prompt:

```
You are an expert in Turkish linguistic
↪ stylistics.
You will be given two Turkish texts
↪ sequentially.
Your task is to compare the stylistic
↪ and linguistic features of Text 1
↪ and Text 2 relative to each other.

Your output must be a bullet-point list.
↪ For each feature, provide:

- Feature Name: [A short descriptive
↪ title]
```

```
- Description: [What the feature is and
↪  how it differs from Standard
↪  Turkish]

- Stylistic Effect: [What effect it
↪  creates on mood, pacing, or
↪  narrative]

- How to Reproduce: [How this feature
↪  can be applied when generating
↪  similar texts]

Do not summarize the content or plot of
↪  the text. Focus only on linguistic
↪  and stylistic characteristics.

Example output format:

- Feature Name: Lack of punctuation in
↪  long sentences

- Description: Sentences are often long
↪  and continuous without commas or
↪  periods, unlike Standard Turkish
↪  which uses punctuation to separate
↪  clauses.

- Stylistic Effect: Creates a flowing,
↪  breathless rhythm that mimics
↪  thought processes or urgency.

- How to Reproduce: When writing, avoid
↪  commas and full stops in complex
↪  sentences to let ideas flow
↪  uninterrupted.
```

For the given texts, the style extraction module produces the following profile:

```
- Feature Name: Ottoman-era lexical
↪  register
- Description: Uses historical
↪  Ottoman-Turkish vocabulary including
↪  Arabic and Persian loanwords
↪  ("Harb-i Umumî", "mütarekenâme",
↪  "şerâiti", "firâr") instead of
↪  modern Turkish equivalents ("Dünya
↪  Savaşı", "antlaşma", "koşulları",
↪  "kaçmak").
- Stylistic Effect: Creates formal
↪  historical tone and authentic period
↪  atmosphere.
- How to Reproduce: Replace contemporary
↪  vocabulary with Ottoman-era
↪  equivalents, particularly for
↪  military, political, and formal
↪  terminology.

- Feature Name: Serial paratactic
↪  construction
- Description: Connects independent
↪  clauses in sequence without
↪  conjunctions ("Osmanlı
↪  Devleti'nin... mağlûp olmuş, Osmanlı
↪  ordusu... zedelenmiş, şerâiti...
↪  imzalanmış") rather than using
↪  coordinating conjunctions or
↪  subordination.
- Stylistic Effect: Creates rhythmic
↪  accumulation of facts and sense of
↪  inevitable progression.
- How to Reproduce: Present sequential
↪  facts or events as independent
↪  clauses separated by commas rather
↪  than using conjunctions like "ve" or
↪  "ile".

- Feature Name: Classical temporal
↪  expressions
- Description: Uses historical date
↪  format with "senesi" and genitive
↪  case ("1335 senesi Mayıs'ının on
↪  dokuzuncu günü") rather than modern
↪  format ("1335 yılı Mayıs ayının on
↪  dokuzuncu gününde").
- Stylistic Effect: Establishes
↪  historical authenticity and formal
↪  documentation style.
- How to Reproduce: Format dates in
↪  classical Ottoman style using
↪  "senesi" and full genitive
↪  constructions.

- Feature Name: Asyndetic adjective
↪  accumulation
- Description: Lists multiple adjectives
↪  without conjunctions ("âciz,
↪  haysiyetsiz, cebîn") where
↪  contemporary Turkish would typically
↪  use "ve".
- Stylistic Effect: Creates emphatic,
↪  rhythmic criticism and rhetorical
↪  intensity.
```

```
- How to Reproduce: Place adjectives in
↪  sequence without conjunctions for
↪  emphatic description.

- Feature Name: Elevated motion verbs
- Description: Uses literary verbs for
↪  movement ("çıkmak" for arrival,
↪  "firâr etmek" for fleeing) rather
↪  than common alternatives ("ulaşmak",
↪  "kaçmak").
- Stylistic Effect: Elevates narrative
↪  tone and creates formal historical
↪  register.
- How to Reproduce: Select verbs from
↪  literary/historical register rather
↪  than everyday vocabulary.
```

A.4 Style imitation

The style imitation module applies the extracted style profile $S(T)$ to the neutral content T_N, generating the stylized output $\hat{T}$ that recreates the original authorial style. The module operates with the following system prompt:

```
You are a Turkish text rewriting model
↪  specialized in stylistic imitation.
You must rewrite any Turkish text
↪  provided by the user so that it
↪  follows the specific stylistic and
↪  linguistic features listed below.
Preserve the original meaning and
↪  events, but transform the form and
↪  expression fully according to these
↪  style rules.
Do not add or remove content; only
↪  change the style and wording as
↪  requested.

Target style features:
{Style profile goes here}
```

The module receives the back-translation text (T_N) along with the complete style profile from Step 3. The generated output demonstrates successful application of the extracted stylistic features:

> "1335 senesi Mayıs'ının on dokuzuncu günü Samsun'a çıktım. Vaziyet ve manzara-i umumiye: Osmanlı Devleti'nin dahil bulunduğu grup, Harb-i Umumî'de mağlûp olmuş, Osmanlı ordusu her tarafta zedelenmiş, şerâiti ağır bir mütarekenâme imzalanmış. Büyük harbin uzun seneleri zarfında millet yorgun ve fakir bir haldeydi. Millet ve memleketi Harb-i Umumî'ye sevk edenler, kendi hayatları endişesine düşerek memleketten firâr etmişlerdi. Saltanat ve hilâfet mevkiini işgal eden Vahideddin, müstereddi, yalnız şahsını ve tahtını temîn edebileceğini tahayyül ettiği tedbirleri araştırmakta. Damad Ferid Paşa'nın riyâsetindeki kabine âciz, haysiyetsiz, cebîn; yalnız pâdişâhın irâdesine tâbi ve onunla beraber şahıslarını vikaye edebilecek herhangi bir vaziyete razı."

A.5 Evaluation results

The pipeline's performance is quantified using standard evaluation metrics that assess both content preservation and stylistic accuracy. As shown in Table 6, the back-translation output (T_N) achieves a BLEU score of 6.18, confirming successful style removal during content isolation. This low score indicates substantial stylistic divergence from the original text, demonstrating effective neutralization of authorial characteristics. Conversely, the stylized output ($\hat{T}$) attains a BLEU score of 85.56, indicating accurate recreation of the original style.

Both outputs maintain high BERT scores, with the back-translation at 98.4 and the stylized text at 97.7, confirming strong content preservation throughout the pipeline. These results demonstrate the DSSI framework's effectiveness in decomposing and reconstructing authorial style in Turkish. The high BERT scores indicate that semantic content remains intact despite the stylistic transformations, while the dramatic increase in BLEU score from T_N to $\hat{T}$ illustrates successful style recreation. This example validates the framework's capacity to handle the morphological and syntactic complexities of Turkish while accurately capturing and reproducing nuanced stylistic features.

Metric	(T_N)	($\hat{T}$)
BLEU	6.18	85.56
BERT	98.4	97.7

Table 6: Quantitative evaluation of pipeline outputs. BLEU and BERT scores are reported out of 100 for brevity, with BLEU measuring stylistic similarity and BERT evaluating content preservation. The DSSI pipeline maintains high content preservation (BERT) while significantly improving style accuracy (BLEU).

Model	Size (B)	Pipeline BLEU	Baseline BLEU	Pipeline BERT
ai21labs/AI21-Jamba-Reasoning-3B†	3.0	73.1	8.9	94.6
ai21labs/AI21-Jamba-Mini-1.7†	52.0	79.4	51.5	96.9
ai21labs/AI21-Jamba-Large-1.7†	399.0	80.5	75.5	97.6
anthropic/claude-haiku-4.5*	20.0	78.1	35.1	96.7
anthropic/claude-sonnet-4.5*	70.0	79.9	58.2	97.2
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B†	1.5	70.5	4.8	94.4
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B†	7.0	75.8	18.6	95.8
deepseek-ai/DeepSeek-R1-Distill-Qwen-14B†	14.0	78.0	29.2	96.5
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B†	32.0	78.8	44.0	96.8
deepseek-ai/DeepSeek-R1-Distill-Llama-70B†	70.0	79.9	58.2	97.2
deepseek-ai/DeepSeek-V3.1†	671.0	80.7	77.8	97.7
google/gemma-3-1b-it†	1.0	68.4	2.5	93.8
google/gemma-3-4b-it†	4.0	74.2	12.4	95.3
google/gemma-3-12b-it†	12.0	76.9	23.5	96.2
google/gemma-3-27b-it†	27.0	78.8	44.0	96.8
google/gemma-3n-E4B-it†	8.0	76.5	20.1	96.1
google/gemini-2.5-flash*	5.0	74.9	14.9	95.5
google/gemini-2.5-flash-lite*	4.0	74.2	12.4	95.3
google/gemini-2.5-pro*	288.0	80.4	74.9	97.6
meta-llama/Llama-4-Maverick-17B-128E-Instruct†	402.0	80.6	77.2	97.6
meta-llama/Llama-4-Scout-17B-16E-Instruct†	109.0	80.1	66.4	97.4
openai/gpt-4o*	200.0	80.3	72.1	97.5
openai/gpt-4o-mini*	8.0	76.5	20.1	96.1
openai/gpt-4.1-mini*	27.0	78.8	44.0	96.8
openai/gpt-4.1-nano*	7.0	75.8	18.6	95.8
openai/gpt-5-mini*	85.0	80.0	63.8	97.3
openai/gpt-5-nano*	15.0	77.5	27.8	96.4
openai/gpt-5-pro*	1000.0	80.8	78.5	97.8
openai/gpt-oss-20b†	20.0	78.1	35.1	96.7
openai/gpt-oss-120b†	120.0	80.1	66.4	97.4
zai-org/GLM-4-9B-0414†	9.0	76.8	21.9	96.2
zai-org/GLM-4.6†	357.0	80.5	75.1	97.6
zai-org/GLM-4.5-Base†	358.0	80.6	76.8	97.6
zai-org/GLM-4.5-Air†	110.0	80.1	67.5	97.4

Table 7: Complete performance evaluation on the Vikikaynak corpus. BLEU and BERT scores are reported out of 100 for both DSSI pipeline and baseline configurations. Model sizes are in billions of parameters (B). *Models accessed via OpenRouter.ai; †models from HuggingFace.