What's in a prompt?
Language models encode literary style in prompt embeddings

Raphaël Sarfati
Cornell University
raphael.sarfati@cornell.edu

Haley Moller
Yale University
haley.moller@yale.edu

Toni J. B. Liu
Cornell University
jl3499@cornell.edu

Nicolas Boullé
Imperial College London
n.boulle@imperial.ac.uk

Christopher Earls
Cornell University
earls@cornell.edu

Abstract

Large language models use high-dimensional latent spaces to encode and process textual information. Much work has investigated how the conceptual content of words translates into geometrical relationships between their vector representations. Fewer studies analyze how the *cumulative* information of an entire prompt becomes condensed into individual embeddings under the action of transformer layers. We use literary pieces to show that information about intangible, rather than factual, aspects of the prompt are contained in deep representations. We observe that short excerpts ($10 - 100$ tokens) from different novels separate in the latent space independently from what next-token prediction they converge towards. Ensembles from books from the same authors are much more entangled than across authors, suggesting that embeddings encode stylistic features. This *geometry of style* may have applications for authorship attribution and literary analysis, but most importantly reveals the sophistication of information processing and compression accomplished by language models.

1 Introduction

"What's in a name?" famously asked Juliet (Shakespeare, ca. 1599) to interrogate the relationship between a concept's multifaceted reality and its shorthand designation in the form of a word.[1]

Four hundred years later, the question finds renewed significance in the context of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Grattafiori et al., 2024), where words are represented as vectors in a high-dimensional latent space (Mikolov et al., 2013a,b). Much research has attempted to elucidate what information these representations, also called 'embeddings', convey, and how this information is encoded. Some fascinating insights have been uncovered in terms of geometrical relationships between concepts (Mikolov et al., 2013a; Park et al., 2024, 2025).

Yet, word-to-vec(tor) embedding is only the first step. For LLMs, an embedding leaves its starting point and is transported, transformer layer after transformer layer, to a new location that will determine next-token prediction. In the process, it loses its original identity and starts accumulating information about all preceding tokens – and the emergent meaning of their sequence (Fig. 1).

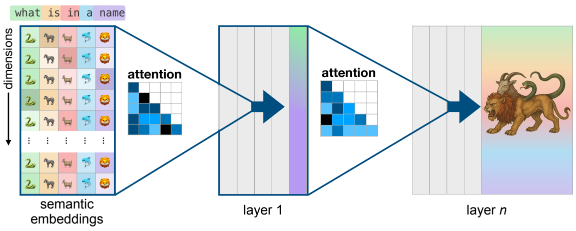

Figure 1: After semantic embedding of the prompt, vectors represent a single word. As the prompt passes through transformer layers, the attention mechanism funnels more and more information about preceding tokens into the last embedding – turning it into a 'chimera' vector, encoding bits of information from all others.

This raises the question: What's in a *prompt*? In other words, what kind of information contained in the sequence of words forming the LLM's input finds itself distilled into deeper embeddings?

Prior work (see also Appendix A) has shown that embeddings can contain global factual information about, e.g., whether the preceding statement is true or false (Marks and Tegmark, 2024), or its relation to space and time (Gurnee and Tegmark, 2024). This is interesting and sensible: factual understanding seems necessary to output compelling prompt continuation.

Here, we find evidence of the presence of more subtle signals. Using short excerpts from various literary works, we show that the embeddings contain implicit information about the origin of the

[1] *"That which we call a rose / By any other word would smell as sweet." Act II, Scene II*

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24059–24068
November 4-9, 2025 ©2025 Association for Computational Linguistics

passage and can be classified with high accuracy.

This study is *not* aimed at merely assessing the performance of LLMs for authorship attribution (Huang et al., 2025), but rather at showing that implicit prompt features like authorship are encoded in deep embeddings (and not early ones).

2 Methods

The code for generating and processing data is available at github.com/rapsar/geometry-of-style.

Overview. We base our analysis on ensembles of short excerpts from various literary works and collect the embeddings *of the last (rightmost) token* after each layer of an LLM. From these vector representations, we apply classifier techniques to evaluate whether excerpts can be linked to their original oeuvres based on a *single*, information-rich embedding. We investigate in particular the influence of context length N (number of tokens in the input passage) and layer depth L (number of transformer layers that the prompt has crossed).

Sources. We use digital versions of literary works obtained from the Project Gutenberg website (`gutenberg.org`). We curate a corpus of 19th and early 20th century anglophone novels for both consistency and diversity of styles, some of them from the same author (Appendix B.1).

Processing. A full novel's text is passed through a model's tokenizer. The sequence of tokens' IDs is then split into chunks of length N tokens, with $N = 8, 16, \ldots, 128$ typically. Importantly, these text chunks do not correspond to any particular syntactic unit and can end with any kind of token (Appendix B.1). The rightmost embeddings $\vec{x}_N(L)$ are collected after each layer L for each chunk.[2]

Models. We use a suite of open-source models from `Hugging Face`. We focus on the 16-layer `Llama-3.2-1B` base model (MetaAI, 2024) for insight and extend to other models in Appendix D.

Classifiers. We use standard supervised classifying techniques to investigate the separability of ensembles of high-dimensional vectors: Support Vector Machine (SVM) probes for binary classification, and Multilayer Perceptron (MLP) probes for multiclass (details in Appendix B.1).

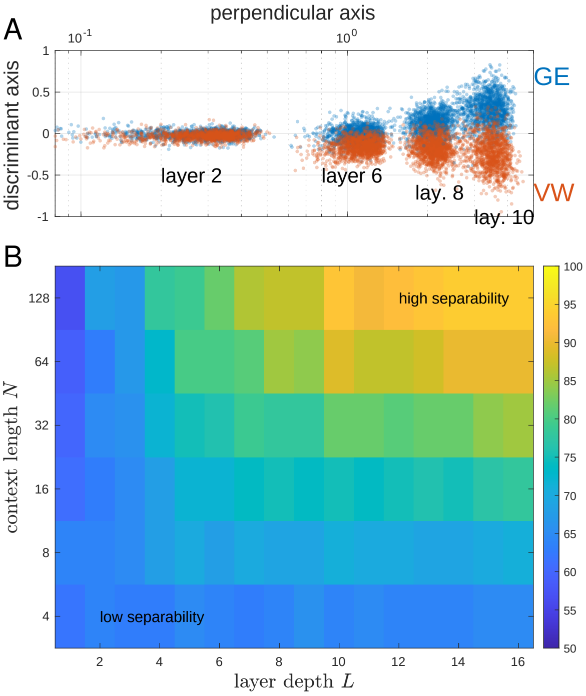

Figure 2: (A) Ensembles of short excerpts ($N = 64$ tokens) from GE and VW separate in the latent space as embeddings travel through successive transformer layers. (B) Linear classifier accuracy (%; average over 10 training and eval runs) to distinguish GE vs VW ensembles as a function N and L.

3 Results

3.1 Embeddings encode authorship

Does a short passage (10-100 words) from a novel contain enough information to be properly attributed after processing by an LLM? We compare excerpts from two novels: George Eliot's (GE) *Silas Marner* and Virginia Woolf's (VW) *Mrs Dalloway*. By training a linear classifier, we examine whether the two ensembles of high-dimensional embeddings in the Llama 3.2 1B model can be separated. They can. We observe in Fig. 2 that as their length N increases and the embeddings travel deeper into the model, the excerpts can be classified with over 90% accuracy. In contrast, when there is not enough context (small N) or not enough attention layers to 'cross-pollinate' information across tokens (small L), each excerpt's last embedding has not absorbed enough contextual information to reflect authorship.

More generally, an MLP probe (3 hidden layers) can distinguish excerpts from several novels with overall 75% accuracy (Fig. 3A). These passages represent small snippets of text from various works, with no consistency in theme or syntax (see

[2] Indeed, the last embedding of the prompt is the one that learns from all preceding tokens thanks to the causal masked attention mechanism.

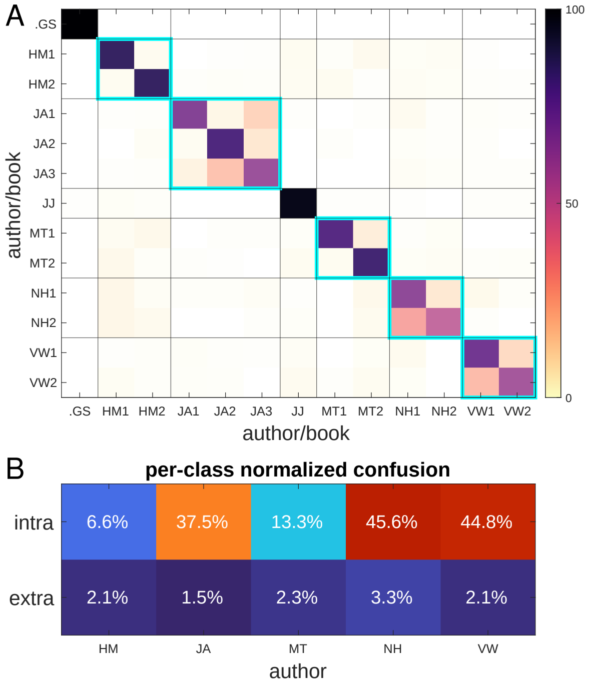

Figure 3: (A) Accuracy (%) of an MLP probe to distinguish passages from 13 different books ($N = 128$, $L = 16$). See Tab. 3 for the list of authors and novels. Cyan squares emphasize novels from the same authors. It is noteworthy that confusion increases between books of the same author, even though they relate to different topics. (B) Results specific to probe confusion for books from the same author (intra) or a different author (extra).

Tab. 4 in Appendix B.1). It could be that they contain enough *factual* information (names, subjects, etc.) to reveal their provenance. However, we also observe a marked increase in classifier confusion across works from the same authors (Fig. 3B). This suggests that the classifier might be relying on patterns of vocabulary and syntax which find themselves encoded in deep embeddings (and not early ones). We refer to these abstract distinctive features as "style" and investigate what exactly is encoded and how.

In Appendix E, we use synthetic textual data to show that these results may not be attributed to model memorization or confounding by topic or vocabulary.

3.2 Stylistic signatures align with large principal components

It's been observed that embeddings and their trajectories tend to form low-dimensional structures. For example, Viswanathan et al. (2025) showed that the intrinsic dimension (ID) of token representations from a given prompt is generally much smaller than that of the ambient space. Sarfati et al. (2025) found, using singular value decomposition, that prompt ensembles stretch along a few directions and diffuse about most of the remaining subspace.

Is the property of "style" contained in the small or large dimensions? Fig. 4 indicates that probe accuracy plateaus at a maximum when keeping about 16 directions along the largest principal components. Interestingly, however, the ID of the ensembles remains under 20 dimensions and doesn't change with increased context N. This suggests that contextual information effectively moves ensembles into different corners of the latent space rather than altering their shape complexity.

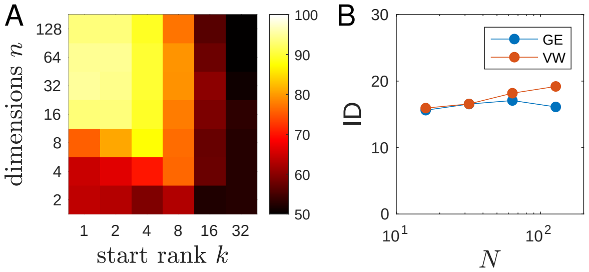

Figure 4: Dimensionality of stylistic features. (A) Probe accuracy (%) for classifying GE vs. VW ensembles projected onto PCA subspaces spanned by $\{\vec{u}_k, \ldots, \vec{u}_{k+n-1}\}$, where $\vec{u}_k$ is the k-th principal component and n is the subspace dimension (B) Intrinsic dimension for embedding ensembles as a function of context length N. ID is calculated using the TwoNN method described in Valeriani et al. (2023).

3.3 Disrupting syntax conserves separability

Is style inferred from syntactic or semantic features? To investigate, we use a shuffling approach introduced in Viswanathan et al. (2025). For each input pseudo-sentence, blocks of B consecutive tokens are rearranged randomly, with $B = 1$ (every token is independent), $B = 4$ (groups of four tokens are kept together), etc. Perhaps surprising, Tab. 1 shows that classifier probes remain accurate for all persistence lengths B. This strongly suggests that the stylistic signature perceived by LLMs might rely not merely on textual structure but rather on lexical content or "distributional information", as also observed in other contexts (Sinha et al., 2021; O'Connor and Andreas, 2021).

3.4 Geometrical relationships across languages

Famously, LLMs latent spaces exhibit alluring geometrical properties, notably the parallelogram structures such as `woman:queen::man:king` (Li et al.,

VW \ GE	$B = 1$	$B = 4$	$B = 32$
$B = 1$	97	100	100
$B = 4$	100	95	98
$B = 32$	100	99	92

Table 1: Linear probe accuracy (%) for various shuffling block size B.

2025). Do similar relationships exist at the ensemble level? Tab. 2 suggests that they do. We consider three French novels and their English translations: Gustave Flaubert's *Madame Bovary*, Victor Hugo's *Ninety-Three*, and Émile Zola's *Germinal*. A probe trained to separate a French ensemble pair keeps its accuracy on the corresponding English pair. Similarly, there is a strong similarity between centroid separations in French and in English. The cosine distance between author A and author B in French and in English is between 0.5 and 0.6, which is far smaller than expected for random vectors ($1 \pm 1/\sqrt{2048}$). This observation seems to generalize point-based geometrical structures to distributed clouds of embeddings.

	F/H	F/Z	H/Z
🇫🇷	79%	63%	65%
🇬🇧	82%	65%	63%

Table 2: Accuracy of a reference linear probe trained to distinguish Flaubert (F) from Hugo (H) in French, when applied to other ensemble pairs. The probe achieves about the same accuracy when applied to the corresponding English-translated ensembles. It performs significantly worse (60%) when applied to unrelated pairs involving Zola (Z).

4 Discussion and future directions

Practicalities. We remark that a by-product of training LLMs is that they inherit a fine perception of stylistic and informational patterns, even from short passages. Perhaps literature scholars will build upon this idea to implement more sophisticated methods to address some long-standing mysteries and controversies: Was Shakespeare a single writer? Could Émile Ajar have been identified as Romain Gary before illegitimately snatching a second Prix Goncourt (Tirvengadum, 1996)?

Geometry of style. As an insightful application, we consider the geometry of style partially uncovered in this study and produce a low-dimensional representation. In Fig. 5, we propose a *map of style* where we place various oeuvres based on the relative proximity of their corresponding embeddings. We discuss and interpret this visualization under the lens of literature analysis in Appendix C.

Interpretability and implications. Beyond practicalities, the main objective of this work is to further understand the information content of LLM embeddings. Many studies have revealed that LLMs construct world models in their latent space, allowing encoding of many features in vector representations, often linearly (Park et al., 2024). Some of these features are easily interpretable while others remain obscure (Bricken et al., 2023; Templeton et al., 2024). We have shown here that intangible aspects of an input prompt, namely stylistic features, are also abstractly encoded in deep representations.

5 Conclusion

We have shown that LLM embeddings representing short (10^2 tokens) literary excerpts encode enough information to identify their origin. Increased confusion between books of the same author suggests that embeddings convey a stylistic signature specific to a given writer. This includes "artificial" writers: we show in Appendix E that text generated by different GPT models (4o and 4.1) can also be accurately differentiated with linear classifiers.

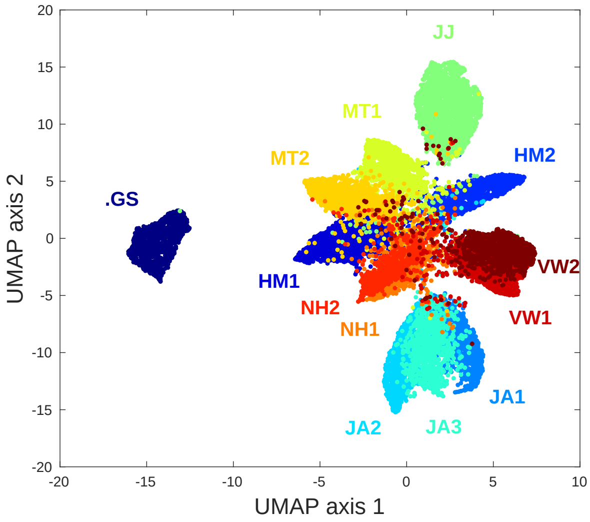

Figure 5: Map of style: low-dimensional visualization of the high-dimensional geometry across books and authors. Text chunks ($N = 128, L = 16$) are UMAP embedded from their 32-dimensional activations extracted at the penultimate layer of the MLP classifier of Fig. 2. We note the substantial overlap between excerpts from the same author, e.g., Austen (**JA1**, **JA2**, **JA3**) or Wolf (**VW1**, **VW2**). More comments in Appendix C.

Limitations

This study is limited to open-source LLMs in the $1-2$ billion parameter range, and a rather small corpus of texts. It also focus primarily on anglophone texts, which tends to constitute most of the model's training set, and then possibly gives the models a finer perception of that language compared to others. When compute in not limiting, testing larger models for probe accuracy will be interesting. In particular, it should reveal whether the accuracy limitation are due to an information bottleneck, or to model limitations.

Ethics Statement

We find that this study complies with the ACL Ethics Policy and the ACM Code of Ethics. We use public domain texts and open-source models for our research, and do our best effort to reference all relevant prior work and acknowledge all contributions. We are not concerned with this study presenting any risk of AI deployment in society. Rather, we anticipate that advances in AI interpretability will contribute to strengthen AI safety guidelines.

Acknowledgments

We thank Jiajun Bao for feedback on the manuscript. This work was supported by the SciAI Center, and funded by the Office of Naval Research (ONR), under Grant Numbers N00014-23-1-2729 and N00014-23-1-2716.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Milad Alshomary, Narutatsu Ri, Marianna Apidianaki, Ajay Patel, Smaranda Muresan, and Kathleen McKeown. 2024. Latent space interpretation for stylistic analysis and explainable authorship attribution. *Preprint*, arXiv:2409.07072.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. *Preprint*, arXiv:2205.10964.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P. Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in Latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412, Miami, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Rebecca M. M. Hicke and David Mimno. 2023. T5 meets tybalt: Author attribution in early modern english drama using large language models. *Preprint*, arXiv:2310.18454.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *Preprint*, arXiv:2408.08946.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. 2025. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. *Preprint*, arXiv:2502.08009.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. 2025. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4).

Qing Lyu, Marianna Apidianaki, and Chris Callison-burch. 2023. Representation of lexical stylistic features in language models' embedding space. In *Proceedings of the 12th Joint Conference on Lexical and*

*Computational Semantics (*SEM 2023)*, pages 370–387, Toronto, Canada. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

MetaAI. 2024. Llama 3.2 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md`. Accessed: 2024-09-25.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, volume 26.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? 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 851–864, Online. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. *Preprint*, arXiv:2406.01506.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Raphaël Sarfati, Toni J. B. Liu, Nicolas Boullé, and Christopher J. Earls. 2025. Lines of thought in large language models. *Preprint*, arXiv:2410.01545.

William Shakespeare. ca. 1599. *The Tragedy of Romeo and Juliet*. Folger Shakespeare Library. Accessed: 2025-04-07.

Adi Simhi and Shaul Markovitch. 2023. Interpreting embedding spaces by conceptualization. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

V. Tirvengadum. 1996. Linguistic fingerprints and literary fraud. *Digital Studies/le Champ Numérique*, 6.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. *Preprint*, arXiv:2302.00294.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. 2025. The geometry of tokens in internal representations of large language models. *Preprint*, arXiv:2501.10573.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Additional background

A.1 LLM internal geometry

LLMs have been found to encode high-level attributes in surprising geometric patterns within their embedding spaces. Recent work supports a linear representation hypothesis, wherein certain abstract concepts correspond to linear directions or subspaces in the latent representation space (Park et al., 2024). For instance, models appear to linearly represent attributes like factual truthfulness, enabling simple probes or even direct manipulation of activations along those concept directions (Marks and Tegmark, 2024). Similarly, categorical semantic relationships can emerge as geometric structures: models represent categories as vertices of a simplex and encode hierarchical relations via approximately orthogonal components (Park et al., 2025). The representational geometry induced by prompts has also been analyzed; different prompting or in-context learning strategies imprint distinct geometric signatures on a model's internal states, highlighting how task framing can alter the organization of latent features (Kirsanov et al., 2025). Moreover, in multilingual settings, LLM embedding spaces can separate language-specific style from language-neutral content along perpendicular axes, suggesting a degree of factorization between surface form and underlying meaning (Chang et al., 2022).

A.2 Authorship attribution

An important high-level attribute of interest is literary style. Authorship attribution and stylistic analysis have served as tests for whether models capture subtle distributional differences beyond topic or semantics. Traditional stylometry relied on carefully engineered linguistic features (e.g. function word frequencies, character n-grams, syntactic patterns), but modern transformer-based LMs can learn such distinctions directly from raw text (Hicke and Mimno, 2023). Recent studies demonstrate that large pretrained models achieve strong performance on author identification. For example, Hicke and Mimno (2023) showed that a fine-tuned T5 model can attribute Early Modern English plays to their likely authors, indicating that LLM representations encode distinctive stylistic signatures. Likewise, GPT-based methods have been applied to Latin prose to verify authorship, with results rivaling traditional stylometric classifiers (Gorovaia et al., 2024). Notably, these analyses also highlight that model judgments can be confounded by semantic content rather than pure style. The challenge of disentangling an author's unique style from the topic of the text is well-recognized in authorship analysis (John et al., 2019; Wegmann et al., 2022; Alshomary et al., 2024).

A.3 Interpretability

Lyu et al. (2023) identified specific latent directions corresponding to concrete stylistic attributes – such as formality and lexical complexity – in a pretrained model's embedding space. Their findings provide evidence that certain stylistic features are encoded along approximately linear axes, making them separable by simple geometric probes. Such results echo the broader concept-vector findings above, but for attributes of writing style. Still, uncovering literary style (encompassing a complex mix of diction, syntax, and narrative voice) may pose an even greater challenge than these relatively focused style elements. Interpretability research has begun to bridge these latent representations with human-interpretable descriptions. Some approaches aim to map regions or dimensions of embedding space to understandable concepts. For example, Simhi and Markovitch (2023) project predefined semantic concepts into a model's embedding space, using them as basis vectors to interpret other embeddings. Another line of work leverages generative LLMs to produce interpretable style representations: Patel et al. (2023) used GPT-3 to annotate millions of sentences with stylistic descriptors and distilled these into a "Linguistically Interpretable Style Embedding" model. The resulting system encodes texts into a 768-dimensional style vector aligned with attributes like formality, tone, and syntax, allowing direct inspection of which dimensions are active for a given text. Together, these efforts underscore both the richness of style information in LLM latent spaces and the complexity of extracting or explaining it.

B Methods

The code for generating and processing data will be made available on a GitHub repository.

B.1 Dataset

In Tab. 3, we present the references to the authors and literary pieces used as the input data to our analysis.

We introduce one text in French (G. Sand's *Jacques*) as a baseline for separability (embeddings from different languages are usually easy to differentiate). The other texts include different anglophone authors with distinctive styles yet from about the same time period, and from different English-speaking countries. This is in order to maintain a certain homogeneity in the style. We also considered only rather large novels (over 500 pages) in order to be able to assemble substantial ensembles.

B.2 Excerpts

The full text of a given novel is tokenized at once, and the resulting sequence of token identifiers is divided into chunks of N consecutive tokens. These non-overlapping chunks represent short text fragments of various forms. In particular, they do not necessarily start or end with a sentence and the last token can be anything: punctuation, word suffix, article, verb, etc. In Tab. 4, we show a few examples of 16-token chunks in their textual form.

B.3 Classifiers

For binary classification, we use a Support Vector Machine with linear kernel, and after dimensionality reduction by PCA to 64 dimensions. The training to validation data split was 70/30.

For multiclass classification, we train a Multilayer Perceptron with penultimate layer of dimension 32, cross-entropy loss and Adam optimizer.

C Map of style

Fig. 5 presents a visualization of style proximity across books and authors. Here we propose an alternative representation and interpret it from a traditional literature analysis perspective.

C.1 Centroid visualization

In order to emphasize proximity between ensembles (rather than text snippets), we propose an alternative representation based on centroid proximity. We calculate the centroids location of each book ensemble and apply multidimensional scaling (MDS) to yield a two-dimensional representation in Fig. 6. The similarity matrix used for MDS is the pair cosine distances between centroids.

C.2 Literary comment

This spatial distribution of Fig. 6 depicts interesting relationships between the narrative techniques and styles of various authors. Virginia Woolf (VW) and James Joyce (JJ), known for pioneering techniques like stream of consciousness and free indirect discourse, cluster apart, reflecting their shared modernist experimentation. Jane Austen's texts (JA), which also employ free indirect discourse but within a more traditional realist framework, form their own distinct grouping. Austen's group is clearly separated both from the modernists and from American authors such as Nathaniel Hawthorne (NH), Herman Melville (HM), and Mark Twain (MT). These American writers, characterized predominantly by narrative realism or romanticism, are grouped centrally and distinctly apart from the experimental modernist approaches of Woolf and Joyce.

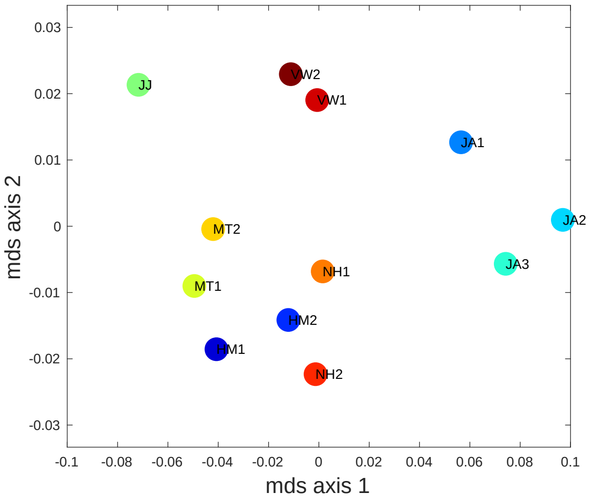

Figure 6: Alternative map of style emphasizing relative distances between ensembles by embedding their shared geometry with multidimensional scaling.

D Generalization to additional models

For generalization, we reproduce the same methodology with three other open-source models released in 2025:

- `gemma-3-1b-pt` from Google (US) (Team et al., 2025)
- `Qwen3-1.7B-Base` from Qwen (China) (Yang et al., 2025)
- `SmolLM2-1.7B` from Hugging Face (France) (Allal et al., 2025)

These models are queried in their "base" form, i.e. not fine-tuned for chat ("instruct"). We use models in the one-billion-parameter range as smaller models are generally more interpretable, and also due to compute constraints.

We find the same patterns of ensemble separability across models, as shown in Fig. 7. In particu-

ID	Author	Novel	Date	Country
GS	George Sand	*Jacques**	1833	France
HM1	Herman Melville	*Moby Dick*	1851	America
HM2	Herman Melville	*Pierre*	1852	America
JA1	Jane Austen	*Emma*	1815	England
JA2	Jane Austen	*Pride and Prejudice*	1813	England
JA3	Jane Austen	*Sense and Sensibility*	1811	England
JJ	James Joyce	*Ulysses*	1922	Ireland
MT1	Mark Twain	*Life on the Mississippi*	1883	America
MT2	Mark Twain	*Roughing It*	1872	America
NH1	Nathaniel Hawthorne	*The House of the Seven Gables*	1851	America
NH2	Nathaniel Hawthorne	*The Scarlet Letter*	1850	America
VW1	Virginia Woolf	*Night and Day*	1919	England
VW2	Virginia Woolf	*The Voyage Out*	1915	England

Table 3: Authors and novels used for analysis. Note that *Jacques* is in French.

15 preceding tokens	last token
`least knew somebody who knew his father and mother? To the peasants of old`	`time`
`off time superstition clung easily round every person or thing that was at all`	`unw`
`crime; especially if he had any reputation for knowledge, or showed any skill`	`in`
`live in a rollicking fashion, and keep a jolly Christmas,`	`Wh`
`shook him, and his limbs were stiff, and his hands clutched the`	`bag`
`nothing strange for people of average culture and experience, but for the villagers near`	`whom`
`road, and lifting more imposing fronts than the rectory, which`	`peeped`
`which seemed to explain things otherwise incredible; but the argumentative Mr.`	`Macey`
`handicraft. All cleverness, whether in the rapid use of that difficult`	`instrument`
`lar or the knife-grinder. No one knew where wandering men had`	`their`
`-weaver, named Silas Marner, worked at his vocation in`	`a`
`of it, and two or three large brick-and-stone homesteads`	`,`
`the outskirts of civilization—inhabited by meagre sheep and thinly-`	`sc`
`certain awe at the mysterious action of the loom, by a pleasant sense of`	`scorn`
`The questionable sound of Silas’s loom, so unlike the natural cheerful`	`tro`

Table 4: Examples of 16-token excerpts from G. Eliot’s *Silas Marner*. Embeddings derived from the last token are the ones collected to form the ensembles of the analysis.

lar, embeddings corresponding to increased context and deep layers are more easily separable.

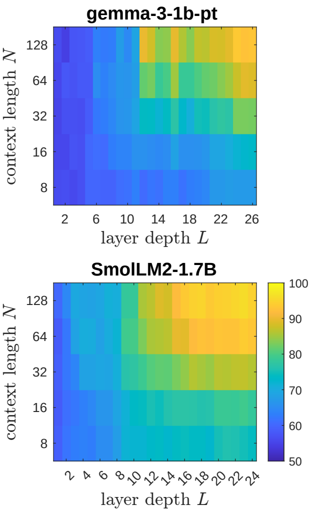

Figure 7: Probe accuracy (%) for GE vs VW in other LLMs (applied on deepest embeddings).

We also observe, unsurprisingly, very good separability at large N and L for the larger `Qwen3-14B-Base` model, this time tested on excerpts from Melville's *Pierre* and Austen's *Emma* pairs (Fig. 8)

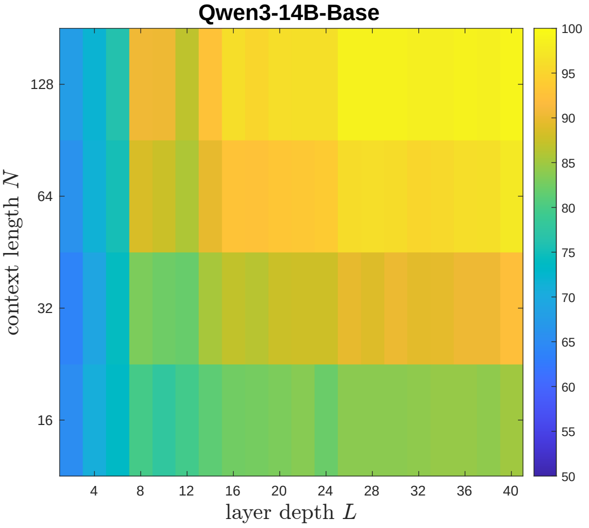

Figure 8: Probe accuracy (%) for HM vs JA in the Qwen3-14B-Base model.

E Controls for authorship probe experiments

E.1 Memorization

One possible concern for the validity of our experiments is that models might have memorized the text of classics during their training. We find this hypothesis unlikely, especially in the case of small 1-10 billion parameter models. Besides, memorization in itself would not necessary cause separation of embeddings. To verify, however, it is easy to input a passage from one the novels used in this study and examine the continuation of base models. Doing so shows that models have not memorized substantial parts of novels. See the Github code to run tests.

E.2 Confounding on topic

Another possible concern is that passages from two novels separate not merely because of the style of their respective authors, but because of the topic of a given novel. Our main experiments suggest that that might not be true, notably because intra-author confusion is increased, even for novels on different topics.

Nonetheless, to investigate rigorously, we created a synthetic dataset by prompting GPT4o and GPT4.1 to write novels in 19th century style, on six different topics (guilt, infidelity, sailing, countryside, marriage, solitude). The different excerpts for each GPT-author were concatenated and passed as chunks through Llama 3.2 1B. Linear classifiers then could successfully separate GPT4o from GPT4.1, even though the data was newly generated and spanned 6 different topics. Results summary below. We indeed find high separability, suggesting that the probes are not differentiating by topic, but rather by style.

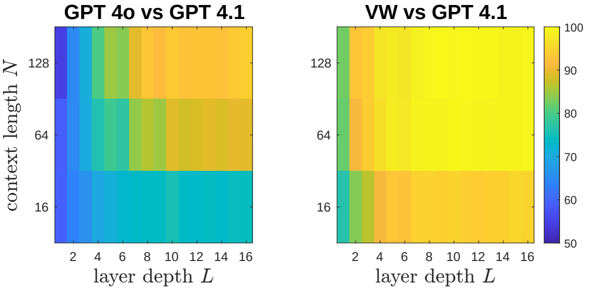

Figure 9: Probe accuracy (%) on Llama 3.2 1B embeddings of text generated by GPT4o vs GPT4.1. We also compared true text from VW to text generated by GPT4.1 with teh following prompt: "novel in the style of V. Woolf on the themes of identity, time and mental health through the eyes of its protagonist".

In the process, this reveals that different LLMs also show different literary style (perhaps the use of "literary" is excessive). This might have applications in AI text detection.