A Deep Dive into Multi-Head Attention and Multi-Aspect Embedding

Maryam Teimouri
TurkuNLP
University of Turku
mtebad@utu.fi

Jenna Kanerva
TurkuNLP
University of Turku
jmnybl@utu.fi

Filip Ginter
TurkuNLP
University of Turku
figint@utu.fi

Abstract

Multi-vector embedding models play an increasingly important role in retrieval-augmented generation, yet their internal behaviour lacks comprehensive analysis. We conduct a systematic, head-level study of the 32-head Semantic Feature Representation (SFR) encoder with the FineWeb corpus containing 10 billion tokens. For a set of 4,000 web documents, we pair head-specific embeddings with GPT-4o topic annotations and analyse the results using t-SNE visualisations, heat maps, and a 32-way logistic probe. The analysis shows that (i) clear semantic separation between heads emerges only at an intermediate layer, (ii) some heads align with specific topics while others capture broader corpus features, and (iii) naive pooling of head outputs can blur these distinctions, leading to frequent topic mismatches. The study offers practical guidance on where to extract embeddings, which heads may be pruned, and how to aggregate them to support more transparent and controllable retrieval pipelines.

1 Introduction

Recent advances in retrieval-augmented models have significantly improved large language models' (LLMs) ability to access and reason over external knowledge (Liévin et al., 2024). A key contributor to this progress is the use of multi-vector and multi-head embedding strategies, which enable richer and more interpretable document representations (Khattab and Zaharia, 2020). These methods have shown strong performance in complex retrieval tasks involving multi-faceted queries. However, critical questions remain about how these embeddings operate internally. In this work, we focus specifically on multi-head representations. While multi-vector approaches also produce multiple embeddings per input, some do so without relying on attention heads. Our analysis is focused on head-based methods, where each representation corresponds to a specific attention head. Do multi-vector models capture more aspects simply because they use larger embedding spaces, or do individual heads learn distinct, complementary features? How much do different heads overlap in what they represent? How far apart are their outputs in the embedding space? Do all heads contribute meaningfully, or could some be pruned without impacting performance? These questions point to a need for more transparent, fine-grained analysis of multi-head embedding behavior. Understanding where to extract embeddings within a model is an important consideration for analyzing their behavior. Previous works ((Zheng et al., 2024), (Besta et al., 2024)) has suggested using representations from the final attention layer, under the assumption that this stage captures the most meaningful structure. However, in the current models' architecture, substantial transformation occurs after the final attention block, which may influence the usefulness or interpretability of these embeddings. Without this consideration, downstream evaluations may obscure or misrepresent the functional roles of individual heads.

In this paper, we investigate the relationship between multi-head attention embeddings and document-level topic structures. To bridge the interpretability gap, we propose a visualization method that maps document-topic alignment to individual head activations, uncovering latent structure within the embedding space. Our work is supported by an automated data pipeline, including topic label annotation of web documents using LLMs. Through a series of visualization experiments and similarity-based evaluations, we examine how alignment with topics varies across heads, how responsive individual heads are to different topics, and how head activity levels influence the resulting document embeddings and their representations. In addition, we

Proceedings of Recent Advances in Natural Language Processing,pages 1263–1270
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-146

conduct a detailed examination of the model's internal structure to identify the most informative stage for embedding extraction, so that the representations we analyze reflect meaningful semantic organization. This process is guided and validated through visualization, allowing us to isolate and study embedding behavior with greater precision. Furthermore, we run comparisons revealing possible mismatches and inconsistencies between external label assignments and internal embedding structures.

2 Related Work

Understanding what pre-trained transformers attend to has become a central topic in NLP interpretability research. One influential study investigates the attention mechanisms in BERT, revealing that many attention heads focus on syntactic roles such as determiners, prepositional objects, and coreference links. These findings suggest that BERT captures rich syntactic structures internally through its attention layers (Clark et al., 2019). Building on this idea, other research has examined the contribution of individual attention heads in multi-head self-attention architectures. It was found that a small subset of highly specialized heads carries most of the model's performance burden. Using a novel pruning approach, the authors showed that a significant portion of heads can be removed with negligible performance drop, highlighting redundancy in attention layers and pointing to opportunities for model compression (Voita et al., 2019). Head pruning can be useful in model quantization, where reducing computational cost without performance loss is a central goal. Transformer quantization research by Bondarenko et al. (Bondarenko et al., 2023) has shown that strong activation outliers often originate from specific attention heads that attempt to perform "no-op" updates by pushing attention scores to extremes. These outliers hinder low-bit quantization. To address this, clipped softmax and gated attention mechanisms are introduced to suppress such behaviors during training. This reduces outlier magnitude and improves quantization compatibility without sacrificing model performance. Beyond language and efficiency, similar head-wise specialization has been observed in other modalities, such as music. In generative music modeling, attention head probing has revealed that individual heads can independently capture distinct musical properties, such as instrument identity or rhythm. This head-wise specialization supports more interpretable and controllable generation, suggesting parallels to the modular roles observed in language models (Koo et al., 2024).

While these studies highlight the role of attention in interpretability, efficiency, and control, recent work also explores how attention can support retrieval-based generation. Retrieval-Augmented Generation (RAG) (Jurafsky and Martin, 2023) combines traditional retrieval techniques with neural generation by first retrieving relevant documents and then conditioning a language model on both the query and the retrieved content. This hybrid framework allows models to access external knowledge beyond their training data, addressing limitations in parametric memory and improving factual accuracy in open-domain tasks. Building on both approaches, Multi-Head RAG (MRAG) (Besta et al., 2024) extends RAG to handle complex queries that require synthesizing information from semantically diverse sources. Unlike standard RAG, which relies on a single embedding vector for retrieval, MRAG constructs a multi-aspect embedding by leveraging the activations from the Transformer's multi-head attention layer, capturing diverse semantic facets of the input. This design utilizes different heads to specialize in distinct semantic aspects, enhancing recall for multi-faceted queries. MRAG has been shown to achieve up to 20% improvements in relevance over baseline methods and integrates seamlessly with existing RAG pipelines and evaluation frameworks such as RAGAS (Tendle et al., 2023). However, important challenges remain. In particular, the interpretability of individual attention head contributions is not well understood, and the mechanisms by which different heads specialize in distinct semantic dimensions are still unclear.

3 Methodology

To begin our investigation, we critically examined a core assumption underlying MRAG: that the multi-vector's embedding models' multiple attention heads capture semantically distinct aspects of document content. While this claim underpins the model's design, it remains unclear whether the observed performance gains stem from meaningful specialization across heads or simply from the larger embedding space and computational capacity.

3.1 Data

Effective evaluation of embedding models requires large-scale, high-quality datasets. The FineWeb corpus (Penedo et al., 2024) offers precisely this: a rich, web-scale dataset that supports both retrieval benchmarking and visualization tasks aimed at uncovering semantic structure in large embedding spaces.

To analyze how individual attention heads respond to different types of content, documents need to be labeled with meaningful categories. To continue this line of investigation using the FineWeb dataset, we applied topic labeling using GPT-4o (OpenAI, 2024). In the FineWeb paper, Appendix F.3 ("Topic Distribution") presents a list of topics and their corresponding distributions. While these topics were originally intended for classification tasks, several of them exhibited semantic overlap (e.g., Math, Formulae, Education and Math, Education, Teaching), while others were overly specific (e.g., Sports, Football, Soccer).

To address this, we merge the original 39 topics and reorganize them into 25 broader categories, aiming to minimize redundancy while ensuring comprehensive coverage across all major themes. Using SFR embeddings (Meng et al., 2024), we generate a heat map of topic similarities with cosine similarity (Salton et al., 1983) as a distance metric. This visualization highlights the semantic relationships between the original topics and guides the merging process. As shown in Figure 1, the selected topics exhibit sufficient differentiation to support meaningful classification and analysis.

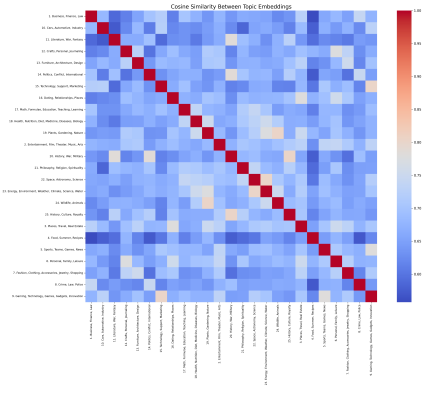


Figure 1: Heatmap of merged topics and their similarities using the SFR embeddings and cosine similarity.

3.2 Model

To continue our investigation, we need to choose a multi-vector embedding-based model. The Semantic Feature Representation (SFR) model (Meng et al., 2024), particularly in its Mistral variant, is designed for dense retrieval tasks. What makes the Mistral-based structure especially relevant for our analysis is its multi-head projection design. This architecture allows multiple ways to extract attention head representations. One approach, stage 1, involves using the embedding vectors directly after the attention layer. This is the stage used in some previous work (Besta et al., 2024); however, since residual connections and normalization layers follow the attention mechanism, important transformations may still occur afterward. To account for this, we define two additional stages for embedding extraction that capture these later processing steps.

The model involves 32 layers, where each layer involves layer normalizations, grouped-query attention (Ainslie et al., 2023), residual connections, as well as up- and down-projection. In the grouped-query attention, the query tensor has the shape `[batch_size, seq_len, 32, head_dim]`, while both the key and value tensors are shaped `[batch_size, seq_len, 8, head_dim]`. To enable attention computation, the key tensor is repeated four times along the head dimension, effectively transforming its shape from `[batch, 8, seq_len, head_dim]` to `[batch, 32, seq_len, head_dim]`. This means that attention heads within a group share the same key and value weight while each have different query weights. This architectural detail is crucial for understanding how information is distributed and reused across heads, and provides a concrete foundation for interpreting the embedding behavior in later stages of our analysis. As part of the attention, the heads are concatenated and transformed through the first projection layer (o_proj), whose output serves as the second point from which we extract embeddings, stage 2.

After the attention, the model includes a residual connection (summing the original layer input with the attention output), layer normalization, up- and down-projecting (projecting from a hidden_size of 4096 to an intermediate_size of 14336 and back), as well as a second residual connection (summing the output of the previous residual with the current projected output). This produces the final model outputs (stage 3), which we slice into 32 parts representing the attention heads. We note that the connection between the actual attention heads and these 32 slices is not necessarily preserved, due to

the two intervening projection layers. To aid in understanding the different stages, Figure 2 provides a schematic overview of the model architecture and indicates where embeddings are extracted at each stage.

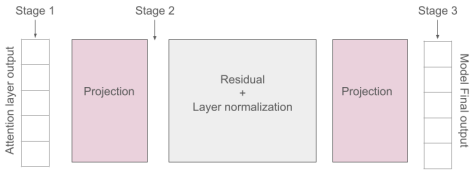


Figure 2: An overview of the model architecture and stages

3.3 Analysis

To begin our analysis, we first visualize the attention heads at stage 1. We use 10,000 documents from the FineWeb Subset: sample-10BT, and SFR embeddings. Each document is processed to produce 32 head-specific embeddings, which we then project into a two-dimensional space using the t-SNE method (van der Maaten and Hinton, 2008). In the resulting plot, each small dot represents one head-specific embedding for a document, yielding 32 dots per document. The dots are color-coded according to their corresponding head index, allowing us to observe clustering patterns and potential distinctions between the roles of different heads. To support exploration of the repeated key tensor, we assigned eight main colors to the 32 heads and varied the shades within each color to correspond to the fourfold repetition. This setup allows us to visually examine whether any patterns related to the repetition are noticeable in the embedding space. As shown in Figure 3a, the 32 attention heads do not form clearly separable clusters. This suggests that several heads may be overlapping or producing similar representations, as indicated by different-colored dots (e.g., purple) appearing on top of areas dominated by another color (e.g., pink). To explore this further, we zoomed in on heads 1–4 (Figure 3b) and observed that some heads appear to be covered by others and are located in close proximity, with only subtle differences in shade within the same main color (red). This reinforces the idea that head-level representations are not yet fully distinguishable at this stage of the model. Although the broader groups appear to cluster well, several individual heads within a group often overlap or lie very close together, suggesting limited differentiation among heads in the same group.

At stage 2 (Figure 4a), we extract the embeddings after the projection layer that follows the multi-head attention layers, where the output is sliced to get the 32 heads. Each head represented by a distinct color consistent with the color scheme of stage 1, are now visible as clearly separated groups. Zooming into heads 1 to 4 (Figure 4b) reveals that these heads are no longer overlapping; instead, there is noticeable space between them. Overall, Figure 4 suggests a transition from stage 1, where some previously observed patterns begin to fade while new ones emerge. The attention heads appear to be becoming more independent in their behavior.

To further investigate the 32 attention heads, we take a different approach by directly slicing the model output to extract the individual head representations (stage 3). In Figure 5a, the heads are color-coded, and the shading indicates the strength of each head's activation: lighter (paler) dots reflect weaker activations, while darker dots indicate stronger ones. Notably, the heads are well-separated and occupy distinct regions in the space, suggesting that each head captures unique aspects of the document representations.

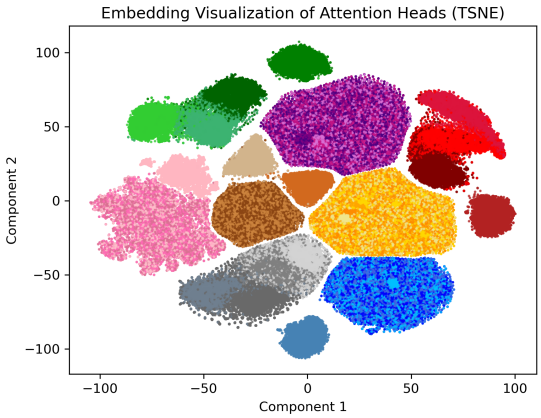
For better visualization, heads 1 to 4 were selected and plotted individually. As shown in Figure 5b, these heads are not scattered randomly; instead, they tend to cluster within distinct regions, indicating consistent behavior.

3.4 Linear Separability

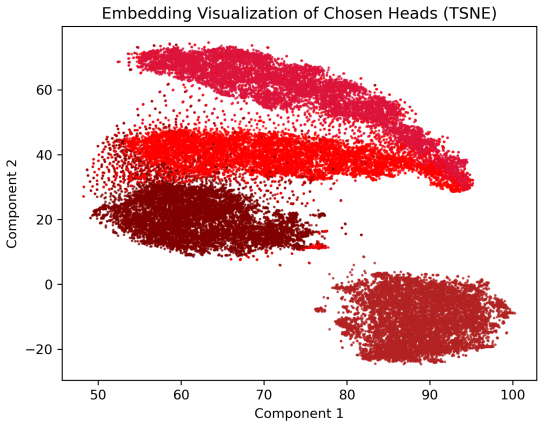
To quantitatively validate the visual structure observed in the t-SNE projection, we performed a multi-class classification analysis using logistic regression. Specifically, we aimed to predict the head index from the vector embeddings to assess the extent to which the heads are linearly separable in this representation space. Head numbers served as class labels, and the associated embedding vectors were used as features. The data was randomly shuffled prior to training. The classifier was trained on 25,600 samples and evaluated on 6,400 test samples. The results, presented in Table 1, confirm that the head-specific embeddings are linearly distinguishable.

3.5 Topic Correlation

To evaluate the alignment between the assigned topics and the SFR model's representations, we de-

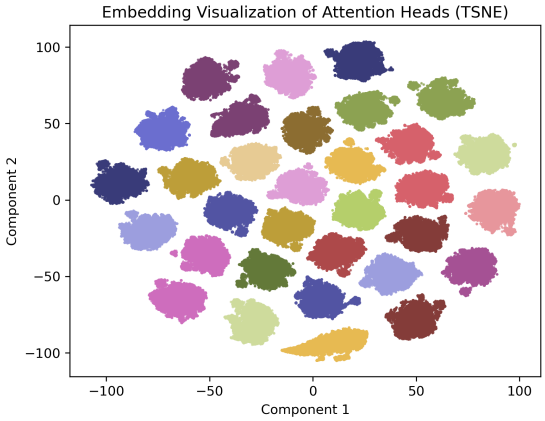


(a) Projection of 32 heads. Not all of the 32 heads are visible; instead, they are grouped into clusters.

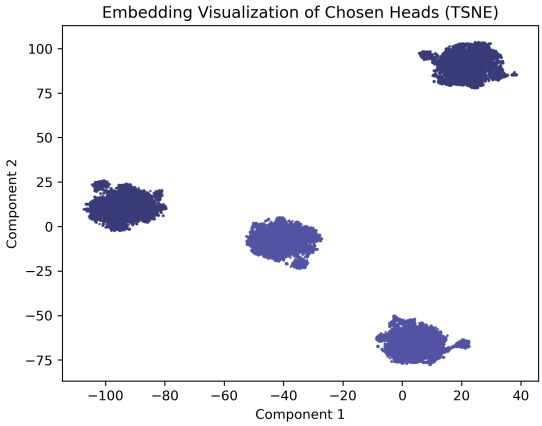


(b) Zoomed into heads 1-4

Figure 3: t-SNE projections with 32 attention heads at stage 1.

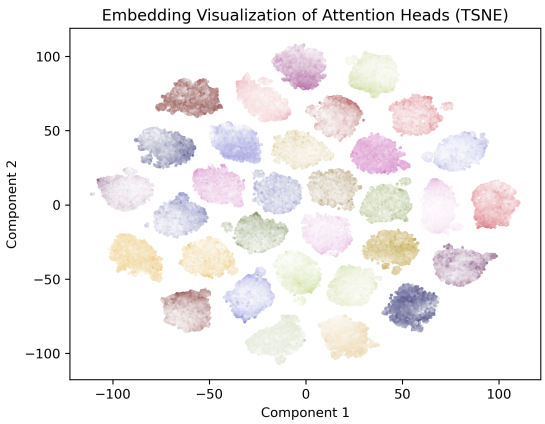


(a) Projection of 32 heads. Grouped clusters have turned into clearly separated individual head clusters.

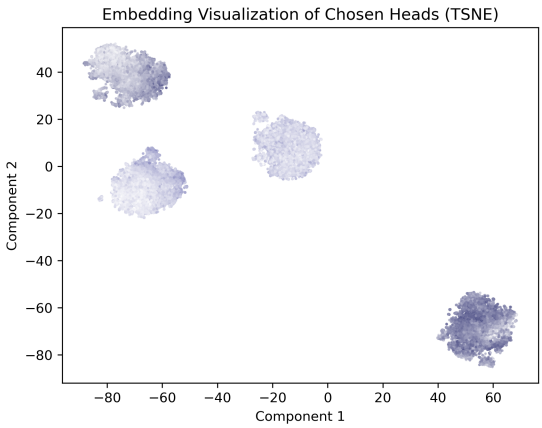


(b) Zoomed into heads 1-4

Figure 4: t-SNE projections with 32 attention heads at stage 2.



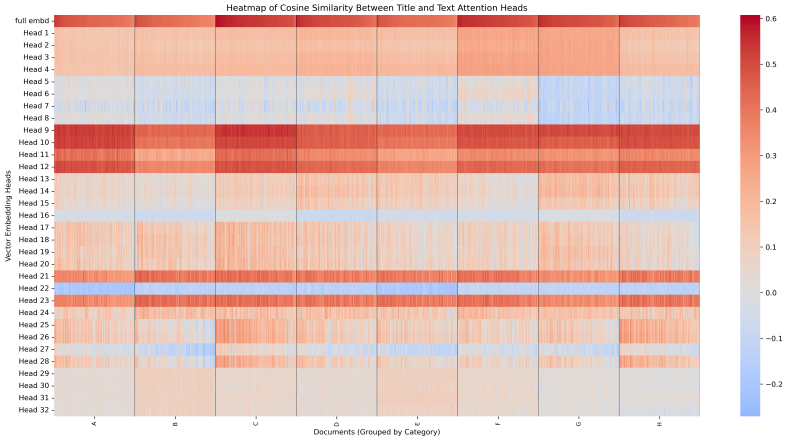
(a) Vector Embedding visualization using t-SNE



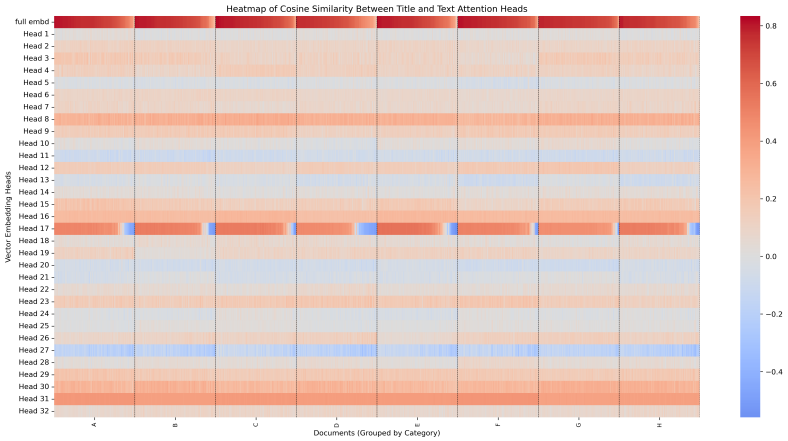
(b) Single heads visualization using t-SNE

Figure 5: t-SNE projections with 32 attention heads at stage 3.

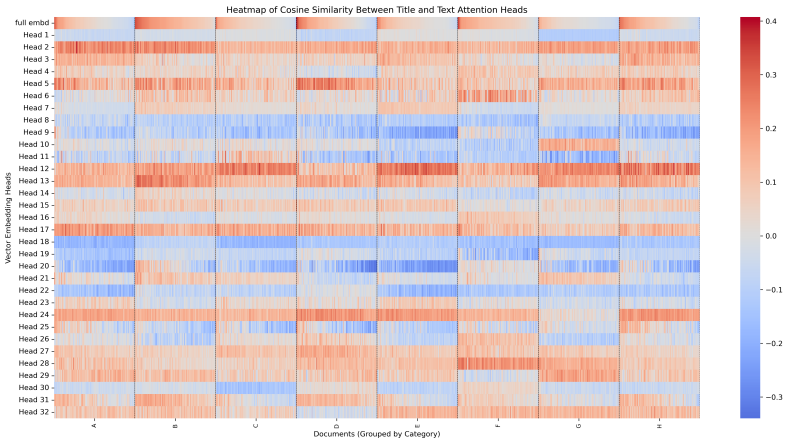
signed a correlation test. The document set, 4,000 articles from the Fineweb's subset sample-10BT, is first classified into our 25 predefined topics using GPT-4o. The same set of documents, along with



(a) Stage 1



(b) Stage 2



(c) Stage 3

Figure 6: Heatmap showing attention head responses by topic. Each topic contains 120 documents. The vertical axis begins with full embeddings at the top, followed by attention heads 1 to 32. The horizontal axis represents document labels A to H, corresponding to the following topics: A. Entertainment, Film, Theater, Music, Arts B. Business, Finance, Law C. Sports, Teams, Games, News D. Gaming, Technology, Games, Gadgets, Innovation E. Personal, Family, Leisure F. Health, Nutrition, Diet, Medicine, Diseases, Biology G. Politics, Conflict, International Affairs H. Places, Travel, Real Estate

the 25 predefined topics, was processed through the SFR model to generate vector embeddings. Cosine similarity was applied to evaluate the accuracy of the topic assignments generated by GPT-4o. The analysis showed that the correct topic appeared as the top-1 match for 30.57% of the documents. When considering the top-5 most similar topics, the match rate increased to 64.82%, and further rose

Stage	Accuracy
Stage 1	0.964
Stage 2	1.000
Stage 3	0.999

Table 1: Accuracy of logistic prob for each stage

to 79.06% when the top-10 matches were taken into account. Although the top-1 accuracy may appear modest, the results suggest that the SFR model captures meaningful topic-related structures in the embedding space. The subsequent heatmaps further explore this relationship by visualizing how different attention heads respond to topic-labeled documents, providing a more detailed view of topic sensitivity across the model's internal representations.

In Figure 6, a heatmap illustrates the similarity between 32 vector embeddings (heads) and 960 documents on 7 topics. The color bar on the right represents the similarity scale, where red indicates higher similarity and blue indicates lower similarity (i.e., greater dissimilarity). The "full embed" row at the top represents the full document embeddings. Within each topic column, documents are sorted based on their similarity between the full document embedding and the corresponding topic embedding. In Figure 6a, which corresponds to stage one, generally sets of four consecutive attention heads (i.e., one group) behave similarly to each other. However, no clear signal is observable across documents, topics, or individual heads. In Figure 6b (stage two), this pattern across sets of four heads disappears, and the heads begin to behave more independently. Still, no strong alignment with document or topic structure is evident. One notable exception is head 17, which follows the sorting pattern of the full embedding row, suggesting the emergence of some meaningful structure. Moving to Figure 6c, attention heads exhibit different attitudes: for instance, head 10 is selective, responding only to the topic 'politics, conflict, international', whereas head 12 responds to a general feature shared across all documents. Head 18 appears to actively avoid one of these shared characteristics. Meanwhile, head 13 seems sparse, reacting independently to individual documents, while head 1 is uniformly smooth, treating all documents similarly. When sorting documents based on their full embedding similarity, head 25 aligns well with this ordering, whereas head 11 does not.

4 Discussion and Conclusion

Our investigation set out to determine whether the advantages of multi–head document embeddings stem from genuine semantic specialization. Given that substantial transformation occurs after the final attention layer, Stage 1 may be quite early to harvest the embeddings. The results provide converging evidence that head-level specialization does exist, but also highlight the importance of the layer from which embeddings are extracted, indicating that meaningful structure may only emerge at certain depths of the model. Somewhat unexpectedly, the strongest topic-wise signal appeared when we directly sliced the final embedding into 32 parts.

The t-SNE projections in Figures 5a and 5b show that the 32 heads carve the space into largely disjoint regions: each head gives rise to a distinct clustering pattern, suggesting that the embeddings they produce capture different structural aspects of the data. The heat-map in Figure 6c further qualifies this observation.

These results suggest that attention heads do not act uniformly or redundantly. Instead, they display specialized, sometimes contrasting behaviors—some being topic-specific, others capturing general or even orthogonal features. This diversity supports the idea that attention heads operate as distinct functional units rather than simply forming a unified embedding vector.

This insight reinforces the value of multi-head architectures for semantic modeling and highlights the potential for more targeted embedding extraction strategies in retrieval-augmented systems.

5 Future Work

Understanding the internal behavior of attention heads reveals their potential to capture diverse semantic dimensions within complex data. While current models implicitly learn to attend to different aspects such as topic or style, this process remains opaque and largely uncontrolled. By making these latent distinctions more interpretable and steerable, we can move toward models that are not only more accurate but also more adaptable, transparent, and capable of being controlled cheaply at inference time. This approach is particularly valuable for complex datasets that contain diverse features such as language, topic, genre, or register, since models trained for specific tasks often overlook these aspects or fail to leverage them effectively. In future work, we aim to address this

by developing benchmarks for multi-aspect embedding models (e.g., SFR, stella (Zhang et al., 2025)) and datasets, enabling us to selectively control model attention—effectively "switching on or off" focus on particular aspects.

Acknowledgments

This research was conducted as part of the EU Horizon project SEUS – Smart European Shipbuilding (Grant Agreement No. 101096224), funded by the European Union. Additional support was provided by the Human Diversity Consortium under the Profi7 program of the Research Council of Finland. Computational resources were provided by CSC – IT Center for Science.

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Maciej Besta, Ales Kubicek, Roman Niggli, Robert Gerstenberger, Lucas Weitzendorf, Mingyuan Chi, Patrick Iff, Joanna Gajda, Piotr Nyczyk, Jürgen Müller, et al. 2024. Multi-head rag: Solving multi-aspect problems with llms. *arXiv preprint arXiv:2406.05085*.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36:75067–75096.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Dan Jurafsky and James H. H. Martin. 2023. Chapter 14: Question answering and summarization. https://web.stanford.edu/~jurafsky/slp3/14.pdf. Draft chapter from *Speech and Language Processing*, 3rd ed.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Junghyun Koo, Gordon Wichern, François G Germain, Sameer Khurana, and Jonathan Le Roux. 2024. Understanding and controlling generative music transformers by probing individual attention heads. In *IEEE ICASSP Satellite Workshop on Explainable Machine Learning for Speech and Audio (XAISA)*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral: Enhance text retrieval with transfer learning. https://www.salesforce.com/blog/sfr-embedding/. Accessed: 2025-04-11.

OpenAI. 2024. Gpt-4o system card. https://arxiv.org/abs/2410.21276. Accessed: 2025-04-11.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Atharva Tendle, Nikhil Kandpal, Marzieh Saeidi, Sumit Bhatia, and Ankur P. Parikh. 2023. Ragas: An evaluation framework for retrieval-augmented generation. https://arxiv.org/abs/2309.15217. ArXiv preprint arXiv:2309.14850.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.